THE BELL SYSTEM TECHNICAL JOURNAL

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

VOLUME III
1924

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

The Bell System Technical Journal

January, 1924

Relays in the Bell System

By S. P. SHACKLETON and H. W. PURCELL

NOTE: Before they can converse people must either be brought together or virtually be brought into one another's presence by the telephone. Any telephone system must establish talking connections between its subscribers, and these connections must be built up, supervised and disconnected when desired. This work is accomplished by the use of relays of various kinds, and the speed and accuracy of service is largely dependent upon them. There are completed daily in the Bell System about 42,000,000 telephone calls. These involve the successful and accurate operation of over one and one-half billion contact connections daily.

Many kinds of relays are employed in the Bell System, varying from the simple electromagnetic drop to the sequence switch, the thermionic vacuum tube and the panel selector. Today a circuit connection between two subscribers served by manual exchanges in a large multi-office district involves about 21 relays. When these subscribers are served by machine switching offices, the number of relays in a local connection may be as great as 146. It not infrequently happens that in setting up a toll connection more than 300 relays are employed.

In the present paper the relay developments leading up to, and making possible the present communication system, are outlined with particular reference to electromagnetic relays. A few typical circuit applications are given with a discussion of the requirements imposed upon relays which influence their design. Several types of relays are illustrated and their distinctive features are described.

The subjects of relay design, manufacture and maintenance and also telegraph relays will be dealt with in future papers.

INTRODUCTION

IN the vast systems of networks which comprise the Bell System one of the most important and varied devices necessary for giving service is the relay. From its use in small numbers in telegraph circuits and as a "drop" in the early magneto switchboard it has come to be numbered by millions and varies in type from the simple electro-magnet which operates a single contact to the vacuum tube and the complete structure which effects an entire series of switching operations.

When a small number of stations is involved in a communicating system complete flexibility of connection may be obtained by means of simple relays controlling a small number of contacts. As the number of stations increases the number of switching operations becomes so great that the use of simple relays which control small numbers of contacts is not economical. The use of power driven selectors and sequence switches and electro-magnetically operated switches for completing a series of switching operations has therefore become necessary.

In present day systems the relay is as essential to a telephone conversation as the transmitter or receiver. Some idea of our dependence on the device may be had from a consideration of the numbers of simple relays involved in a typical connection. A circuit established between two subscribers served by manual exchanges in a multi-office district will involve 21 relays. When these subscribers are served by machine switching offices the number of relays

Fig. 1—Relays in a local manual office

involved in a local connection may be 146. When toll connections are involved even greater dependence is placed on relays to render service. A New York-San Francisco connection requires over 200 relays and very frequently connections are established which require more than 300 relays. The majority of these relays are normally available for doing their bit to provide telephone service to any one of a large number of subscribers. As a matter of fact, approximately 90 per cent of the millions of relays in the Bell System today are available for and may be called upon to serve any subscriber or user of the telephone.

A typical manual office serving 10,000 lines would have from 40,000 to 65,000 relays and their total combined pull if applied at one point would be sufficient to lift ten tons. In the larger machine switching offices there may be as many as 140,000 relays which require in some instances power plants capable of handling peak loads of 4,000 amperes at 48 volts.

Referring to Fig. 1, the space required for mounting some of the relays in an office will be seen. This is a picture taken in one of the New York offices which has over 60,000 relays and the racks shown contain about 22,000 of these. The covers have been removed from a number of the relays in the foreground. Instead of grouping the relays compactly as in a manual office it is the practice in machine switching offices to mount them in close association with the related apparatus units. This is illustrated by the photograph of sender circuit relays shown in Fig. 2.

Invention of the Electromagnet

Prior to 1820, the electro-magnetic structure, now known as a relay, was an impossibility because the scientific facts on which it is based had not been discovered. In the winter of that year, Oersted of Copenhagen established that a mechanical effect could be produced on a magnetized needle by a current of electricity. Oersted discovered that a magnetic needle would be deflected from its normal position when held parallel to a wire conveying an electrical current and that the deflection would be to the right or left, according to the direction of current flow. This discovery aroused such interest among scientists and philosophers that the best minds in Europe were engaged in speculation and experiment, so that further discoveries of great importance followed rapidly. Arago in Paris and Davy in London, working independently, soon observed that, if an electric current passed through a wire of copper or any other material, the wire had the power of inducing permanent magnetism in steel needles.

Oersted's discovery suggested to Schweiger that the mechanical effect on the magnetic needle would be increased if the current were made to pass several times around the needle. He made a coil, elliptical in shape, of insulated wire and suspended the magnetic

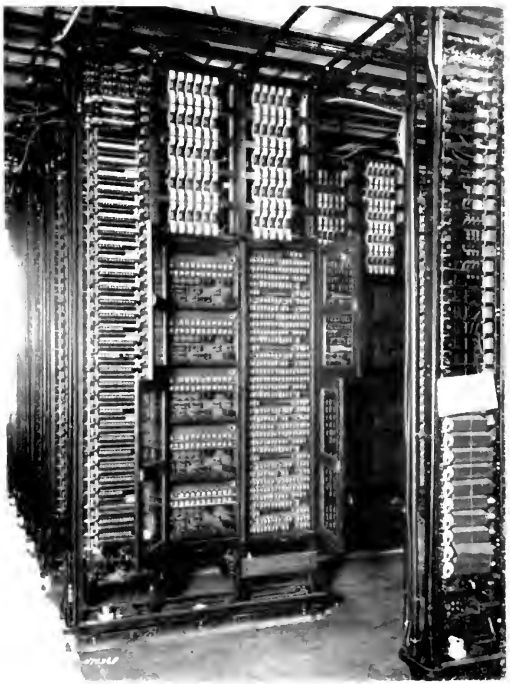


Fig. 2—Sender relay cabinet in machine switching office

needle within it. When current passed through the coil, the result was as he anticipated, and the device became known as Schweiger's multiplier.

Ampere, the brilliant French scientist, in seeking an explanation for Oersted's discovery, evolved an ingenious theory of the relation

between electricity and magnetism. According to this theory, all magnetic phenomena result from the attraction or repulsion of electric currents supposed to exist in iron at right angles to the length of the bar, and all the phenomena of magnetism and electro-magnetism are thus referred to one principle the action of electrical currents on one another. Among other things, he proposed a plan for the application of electro-magnetism to a system for transmitting intelligence. This system was to operate by the deflection of a number of needles at the receiving station by currents transmitted through long wires. By completing a circuit the needle was to be deflected and was to return to normal under the influence of the attraction of the earth when the circuit was opened. This proposed system of Ampere's was never reduced to practice, however.

All these discoveries and results were prior to 1823, and they resulted in the development of needle telegraph systems, which were at one time employed extensively in Europe. These systems utilized a coil of wire around a magnetic needle pivoted in the center and with a pointer attached to the needle, which was suspended over a dial. Deflections to the right or left signified letters and were produced by sending pulsations of one polarity, or alternations of both, as was required.

In 1824, Sturgeon, an Englishman, discovered that, if a current of electricity flows in a coil of wire surrounding a bar of annealed iron, the latter becomes a magnet, and when the current ceases, the iron loses its magnetism. Sturgeon bent an iron rod into the form of a horse-shoe and wound a coil of copper wire around it loosely, with wide intervals between the turns to prevent them from touching each other. Through this coil, he transmitted a current. The iron under the influence of this coil became magnetic and thus, the first electro-magnetic magnet, now known simply as the electro-magnet, was produced. This discovery of Sturgeon's is of great interest to the telephone and telegraph engineer, because it was the direct step which made the invention of the electro-magnetic relay possible.

In 1828, Henry, in America, after repeating the experiments of Oersted, Ampere, Sturgeon, and others and investigating the laws of the development of magnetism in soft iron by means of electrical currents, designed the most powerful electro-magnet that had ever been made. This he accomplished by associating Schweiger's multiplier with Oersted's magnet. For this purpose he wound 35 feet of silk insulated wire around a bent iron bar so as to cover its whole length with several thicknesses of wire.

First Telegraph Relays

In 1824, Morse utilized the electro-magnetic phenomenon, revealed by Sturgeon, and produced a telegraph system which was destined to be the basis of all modern systems of communication. The attenuation of the current from the sending to the receiving end of the circuit had limited the satisfactory transmission of signals. Morse overcame this difficulty by constructing an electro-magnet which would repeat or "relay" the transmitted signals to another circuit having an independent source of energy. The first electro-magnet or relay designed by Morse was a cumbersome structure weighing about 300 pounds, but it exerted a tremendous influence on the art of communication as it served as a stimulus for the development of the complex systems of the present day. When this relay was redesigned its weight was reduced to 70 lbs., but as the laws of electro-magnetism became more generally understood and new materials became available, such great changes occurred that the present telegraph relay weighs about 3 pounds, and one of the modern telephone relays of latest design weighs but 3.6 ounces.

The General Problem

The needs of the present day telephone and telegraph system have produced a multitude of devices but none of them is of greater importance than the relay, as it affords the means whereby the engineer may put ideas into practice. When the limitations of available relays prevent the satisfactory solution of a problem, requirements for new relays are outlined and their development is undertaken if a survey indicates that the advantages to be obtained warrant the expense.

This does not mean that compromises are not made in the matter of using standard designs, for in some instances, it would not be economical to design a new type. Frequently, a relay is required to meet certain conditions in the plant for which the demand will be comparatively limited, and it is obviously uneconomical to spend time and money developing a new type provided a standard structure can be adapted to meet the requirements with sufficient precision.

Just as the art of defense in warfare has matched the art of offense, the art of relay design has kept pace with the demands of the circuit engineer. Relays are now required to operate on direct, and pulsating current, and also on alternating current throughout the entire range of frequencies which are used in communication. There are fast relays, slow relays, polarized, high impedance, low impedance

and so on. Consequently, a relay designed for one purpose may be wholly unfit for any other use. On this account, as the telephone art has grown, new conditions and new requirements have resulted in the development and manufacture of a large number of relays. At first, this undoubtedly followed previously established precedents, so that new forms were brought into existence which fulfilled immediate needs, but did not receive much consideration from the standpoint of economy, standardization or consistency of design.

At the present time, the Western Electric Company manufactures for the Bell System about 100 types of simple electro-magnetic relays. These types are subdivided into about 3,500 kinds, which differ in minor ways, such as windings and contact arrangements. In 1921, the Western Electric Company produced over 4,800,000 of these relays. These figures serve to indicate the economic importance of the relay in the present day system but do not give any adequate conception of the dependence of the communication network on relays of all types.

From a design standpoint it is possible, as has been pointed out, to attain practically any desired result in an electric circuit, subject of course to certain limitations as to time, and provided no limitations as to economic application are to be met. The methods and means for securing the desired operations involve the use of relays of various types and designs, and may lead to new developments which are obviously not economical. The relay may be called upon to perform a single function, necessitating the opening or closing of a single contact, or it may be required to effect a complicated series of transfers or circuit changes. Its operation may necessitate an accurate synchronizing with other circuit operations involving a time lag in its operation or release, and other requirements as to impedance, power, etc., may be imposed. It frequently happens that the conditions imposed by circuit requirements necessitate a choice between new features of relay design and a complication of the circuit to overcome limitations in existing types of relays.

The economic considerations which govern the final application of circuits in the telephone system are, to a large extent, dependent on the costs and performances of the various types of apparatus, particularly the relays. Frequently, there may be a number of possible methods of accomplishing a given result in an electric circuit and the most economical method is, of course, desired. This does not necessarily indicate the least number of relays or the cheapest but rather the most economical combination, taking into account reliability of circuit operation and its effect on service, cost of equip-

ment and cost of operation and maintenance. The more complicated circuit or the one necessitating additional equipment may be sufficiently more reliable to justify its use.

In considering the application of relays in any telephone circuit, a given problem is usually presented and the various possible methods of accomplishing the desired end are considered. These methods may involve combinations of relays or of relay parts which do not exist and may even involve combinations which are entirely impractical or uneconomical of application. Any simplifications which may be effected are considered and in case the design of any apparatus may effect appreciable savings in the circuit or otherwise appear justifiable, this may be undertaken. Such requirements on relay design are, of course, subordinated to any general design considerations, such as relay structure, etc., which may be governing from the standpoint of the economic production of the relays themselves.

A few considerations which influence the selection of relays and which are very closely associated with the fundamental relay design may be considered from the standpoint of their effect on telephone circuits and their application in the field. It would, of course, be impossible within the scope of this paper to describe all the relay applications in modern telephone practice, but a discussion of the relay requirements for a few typical cases will serve to illustrate the principles involved. While the first relays used in the telephone system were telegraph relays adapted for use in signaling, the vast majority of relays now in use in the Bell System are designed primarily for telephone circuits. The requirements are usually quite different, particularly as regards the energy available for operation, the speed of operation and reliability of contacts and in most cases the cost.

Early Telephone Relays

In the first telephone switchboard for commercial service which was installed in 1878, the electro-magnetic devices consisted of a telegraph relay and an annunciator for each subscriber's line, and a call bell common to all lines. Of these three the telegraph relay was the largest and most costly, so the desirability of reducing the number required and of providing a smaller and cheaper apparatus unit was apparent. Changes were soon made in the magneto system that removed the relay from the subscriber's line and associated each relay with a group of lines for supervisory purposes. In the early switchboard, patented in 1879, from which the standard switchboard was developed, a modified telegraph relay appeared as a clearing out

relay to control a clearing out drop. This modified telegraph relay is shown in Fig. 3 and is of particular interest as representing the first step in the development of the telephone relay.

In the magneto systems the indicator or drop was of the first importance, so its development was rapid. It was finally arranged in one extensively used system with two coils, which were known as

Fig. 3—Early telegraph relay used as telephone drop

the line coil and the restoring coil. The magneto current from the subscriber's station energized the line coil to drop a shutter which was restored through the agency of the restoring coil when the operator inserted a plug in the associated jack. The early development of the drop undoubtedly influenced the forms of relay structures which were developed a little later. The analogy between the line and restoring coils of the magneto system and the line and cutoff relays of the common battery system is very close. In the latter the current over the subscriber's loop energizes the line relay which lights the line lamp. The insertion of the plug in the subscriber's jack energizes the cutoff relay which opens the circuit through the line relay and thus extinguishes the light. In addition, the line and cutoff relays are assembled on a common mounting plate, forming an apparatus unit, although they are not parts of the same structure as were the corresponding coils of the drop.

The early telephone relays resembled more closely in construction and form the early drop than they did the telegraph relay, although the influence of design work on the telegraph relay appears in the development of later types of telephone relays. The early

telephone relay shown in Fig. 4 has little relation with the modified telegraph relay of Fig. 3. It is much smaller and lighter than the first relay and, in addition, there is a distinctive structural change in that the armature is suspended by a reed hinge.

At this time the limiting conditions controlling the operation of either telephone circuits or the apparatus in them had not been

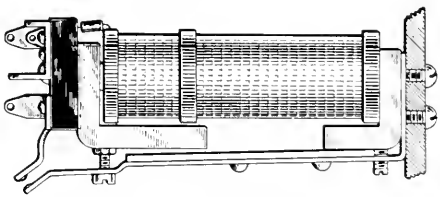


Fig. 4—Early telephone drop relay

established with much precision, so that the requirement for a relay was, roughly, that it should do the work required and any arrangement more sensitive and reliable than a previous arrangement was an improvement. The principle of the reed hinge for an armature support was sound, in that it provided for a good magnetic circuit and an easy means for close air gap adjustment and it is now used extensively with relays of the latest design.

Line, Cutoff and Supervisory Relays

When the common battery system was developed, however, it was found that the reed hinge relay was not capable of meeting the additional requirements imposed by the new system. The common battery cord circuit originally suggested is shown in Fig. 5. It is apparent that the relay shown in the diagram must indicate positively to the operator the position of the switch hook in the substation set and must respond to the motion of the switch hook if the subscriber moves the hook up and down to interrupt the circuit. In addition, as this relay is in the transmission circuit it must not introduce objectionable transmission losses. The reed hinge relay could not meet these additional requirements, and accordingly a new relay was designed especially for the common battery system that was the most important single factor in making the new system possible. In order to obtain an armature that would respond quickly to any change in the holding magnetic force all forms of support for the armature were rejected. The relay developed is shown in Fig. 6

and was first used in the common battery board installed in Worcester, Mass., in 1896.

This relay consists of a tubular magnet with an iron disc armature in the form of a truncated cone. This disc is brought to an edge at its periphery and rests in an annular groove in the cap. When the armature operates, it closes against an insulated contact stud

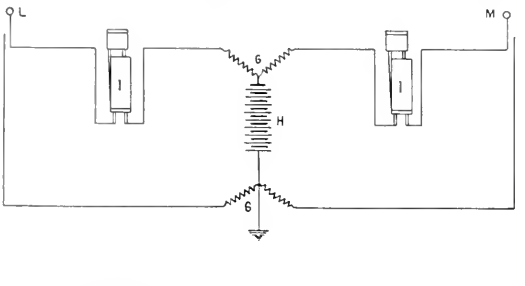


Fig. 5—Early common battery cord circuit

projecting from the core and when released drops away from the core by gravity and rests against a stud projecting from the end of the cap which provides the adjusting means for regulating the armature air gap. As the contacts of this relay were enclosed in the case, they were protected from dust and this arrangement proved so desir-

Fig. 6—Early line relay

able that it has been an accepted feature of nearly all relay designs that have followed. This arrangement had the disadvantage, however, of not providing a means for determining the value of the armature air gap or the contact separation. This condition was improved in the next design which is shown in Fig. 7 by making the cap longer and associating the armature with the magnet structure

instead of the cap. The knife edge armature hinge and the gravity control of the free armature were the fundamental principles retained in the new design.

Fig. 7 Early supervisory relay

The tubular shield for the return magnetic path was abandoned for a return pole piece which provided a means for mounting both the armature and the air gap adjusting screw. The cover protected the contacts from dust but it was soon found that magnetic interference between adjacent relays was responsible for both faults in

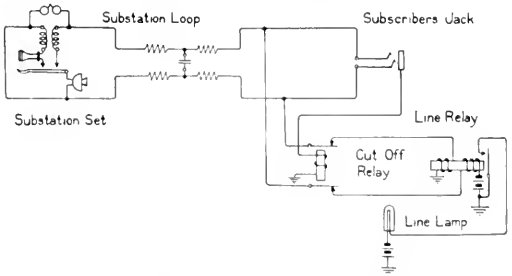


Fig. 8 Line and cutoff relay circuit

operation and crosstalk, so the cover had to meet the additional requirement of being an effective shield. This was eventually accomplished to best advantage by making the cover of copper.

As has been shown two general types of structures were now available for common battery system relays. In one, the armature was

suspended by a reed under tension to provide a restoring force. In the other, which was more sensitive but less capable of carrying a heavy spring load, the armature operated in a knife edge hinge and the restoring force was gravity.

Each subscriber's line required a line relay for lighting a lamp when the substation receiver was removed from the hook, a cutoff relay for removing the line relay from the circuit when the operator responded, and a supervisory relay for controlling lamp signals to inform the

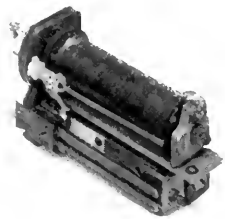


Fig. 9—Line and cutoff relays

operator whether the substation switch hook contacts were open or closed. The circuit arrangement for the line and cutoff relays is shown schematically in Fig. 8.

The rapid extension of telephone service necessitated establishing standards of excellence, and definite requirements for apparatus units were gradually formulated. At first, the available relays were adapted as closely as possible to existing conditions, but as requirements became definitely established, relays were designed specifically to meet them and careful consideration was given to manufacturing costs, mounting space, maintenance expense and all other factors of economic importance. By 1910 several million of the line and cutoff relays shown in Fig. 9 and the supervisory relay of Fig. 7 were in service.

The cutoff relay armature was of the reed hinge type, while both the line and supervisory relays were assembled with the more sensitive knife edge armature. The line relay was eventually wound with 12,000 minimum turns to a resistance of 2000 ohms ± 5 per cent. and after considerable service experience requirements were formulated for a line relay which would be a satisfactory substitute. These requirements were as follows:

(1) Battery potential, 20–28 volts.

(2) Maximum line resistance, including subscriber's station, 1000 ohms.

(3) Resistance across line to represent maximum insulation leakage, 10,000 ohms.

(4) Winding of relay 12,000 turns, 2000 ohms ± 5 per cent.

(5) Minimum operating ampere turns =

$$\frac{\text{Turns} \times \text{Minimum Voltage}}{\text{Maximum Resistance}} = \frac{12{,}000 \times 20}{1000 + 2100} = 77.4.$$

(6) Maximum releasing ampere turns =

$$\frac{\text{Turns} \times \text{Maximum Voltage}}{\text{Leak Resistance} + \text{Minimum Relay Resistance}} = \frac{12{,}000 \times 28}{10{,}000 + 1900} = 28.2.$$

Reference to the circuit will show that the line relay must release on a low resistance loop in case the subscriber flashes the line lamp to attract the operator's attention. Due to residual magnetic effects, a relay does not release after operation on short loops over which the operating current is high as quickly as it does after operating on long loops, with a lower current in the winding. It is, therefore, necessary to specify the maximum ampere turns the line relay may receive and adjust it to release immediately afterward with the maximum leak across the line.

(7) Maximum ampere turns $= \dfrac{12{,}000 \times 28}{1900} = 176.8.$

In addition:

(a) The relay must close one set of contacts which controls a signal lamp as shown in the circuit.

(b) Contacts must carry the energy for lighting the lamp without undue sparking, sticking or wear.

(c) The relay must operate reliably on 77.4 ampere turns.

(d) The relay must release on 28.2 ampere turns immediately after operating on 176.8 ampere turns.

(e) As there is a constant potential between windings, the coil must be protected from corrosion, so the materials chosen for the construction of the relay must not contain substances which tend to encourage or assist electrolytic action.

It was found that the line relay introduced high maintenance charges because of the knife edge armature hinge and the close adjustment required. The armature being comparatively light in weight,

a slight amount of dust or corrosion in the armature slot frequently made the contact resistance in the slot so high that the line lamp would fail to light.

The supervisory relay in the cord circuit also introduces maintenance charges for the same reason. It had to meet the same requirements as the line relay but, in addition, it required a crosstalk proof cover that would also adequately protect the contacts from dust.

The solution of this problem was very difficult because the only means of obtaining relays of increased sensitivity or greater operating range was to make mechanical refinements, which would increase manufacturing costs quite out of proportion to the advantages obtained, to discover new magnetic materials of higher permeability at low flux densities and with a lower remanence characteristic or to develop an entirely new relay structure. To obtain any advantage from the development of a new structure built from the same magnetic materials, it would be necessary to design it in such a manner as would enable the engineer to obtain the proportion of copper and iron required for maximum efficiency, greatest economy, extreme sensitivity, maximum operating load or any other specific requirement which was the controlling factor of a particular design. Analytical studies had shown that smaller relays with less iron could be substituted for those in use but such a change could not be made without increasing manufacturing costs because a reduction in core diameters would increase breakage during manufacture as well as entail a greater cost of handling the smaller structures.

The Flat Type Relay

An analysis of the manufacturing costs had shown labor costs to be greater than material costs in the production of relays so that any changes which would result in large savings would have to be of such a nature as to reduce labor charges. This could be accomplished only by changing production methods which had already been established with reference to greatest economy in manufacture considering the volume of production. The demand for relays, however, was increasing steadily and it was evident that with increased production the prevailing manufacturing methods would not continue to be economical. With other pieces of apparatus manufactured in large quantities, it had been found that production costs could be reduced to a minimum by designing a unit which could be assembled from interchangeable parts stamped out by a punch press and formed in bending fixtures to the desired shapes. To accomplish this for relays, it was first necessary to conceive of an

entirely new type of structure composed of parts that could be made easily by the punch press method, and it was then necessary to determine the modifications which must be made in such a structure, in order that the proportions of copper and iron required for any one of a number of design conditions might be obtained.

The design of a punched type relay was first attempted in an effort to find a better relay than the line relay, and with the intention,

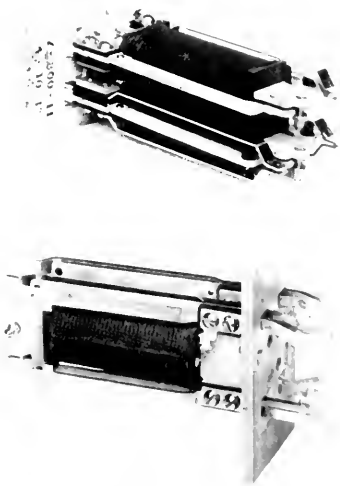


Fig. 10—First punched frame relay

if the design were successful, of employing the same structure with a different winding as a substitute for the cutoff relay. It will be remembered that the line relay had a gravity type armature, whereas the cutoff relay had a reed-hinged armature so the effort to replace two relays of different construction by a single structure was the beginning of an attempt to standardize a type of relay structure which could be used universally.

After some years of development work, a commercial design was completed and punched-type relays were produced as substitutes for the line and cutoff relays. The structures were exactly alike,

the relays differing only in windings and in the number of contact-carrying springs with which they were equipped. The development of these relays resulted in a price reduction for the line and cutoff relay unit of about 25 per cent, and a reduction in the mounting space occupied of 40 per cent. The flat core and the manner of suspending the armature on a reed hinge, in order to present the armature to the pole face, were the distinctive features of the new

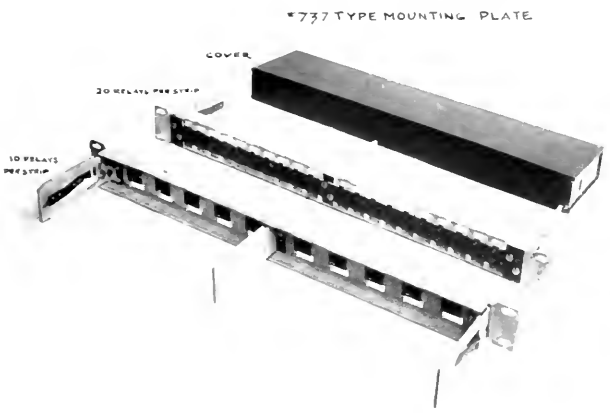


Fig. 11 Mounting plate for strip of punched frame relays

relay structure, as will be seen by referring to Fig. 10, in which the line and cutoff relays may be distinguished because they are equipped with a single pair and a double pair of contacts, respectively. The method of mounting the relays and protecting a strip of 20 with a common dust cover is shown in Fig. 11 from which it will be observed that the mounting plate, all the mounting details and the cover, are products of the punch press.

When it was seen that the development of the new line and cutoff relays was proceeding favorably, development work was also begun on a similar punched-type substitute for the round core supervisory relay which has previously been described. It was known that the quantity of iron in the supervisory relay was greatly in excess of the amount required, as the core flux density was far below saturation when the relay operated over the longer substation loops and

the magnetizing ampere turns were reduced by the high resistance of the loop. Advantage was taken of silicon steel, a new material at that time, which had a higher permeability than Norway iron and less pronounced residual magnetic effects, after saturation. In addition, it had greater tensile strength and, since the new type relay core was rectangular in shape and therefore had the stiffness of a beam, it was possible to make a core of silicon steel of such small cross section that the flux density was much higher with a small

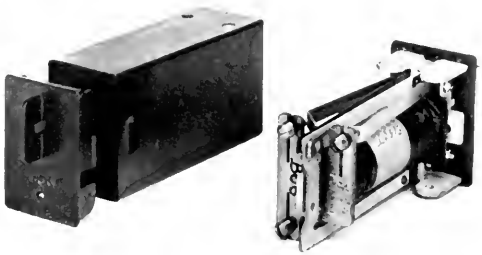


Fig. 12—"B" type relay

magnetizing force than it would be with a Norway iron core of the minimum cross section necessary for structural strength. A supervisory relay was, therefore, produced which was similar in construction to the line and cutoff relays and occupied the same mounting space. It was necessary to develop a dust protecting cover for this new relay which was also cross-talk-proof, in order to prevent the reproduction of conversation by mutual induction between adjacent relays. The design of this relay was such that spring tensions and contact adjustments were controlled by screws mounted in a brass plate at the front of the relay. The increased sensitivity of this relay over that of the round core type permitted the limits for substation loop resistance to be increased from 750 to 1,000 ohms, and the combined resistance of the windings to be reduced from 12 to 9.1 ohms, which decreased the transmission loss in the relay about 30 per cent. In addition, this new relay was superior in flashing ability and also released on a higher number of ampere turns. The

mounting space was reduced 25 per cent. Large savings also resulted from a reduction in maintenance costs from approximately $5.00 per switchboard position per year to a negligible amount. The new relay is shown in Fig. 12, which shows the adjusting screws in the plate

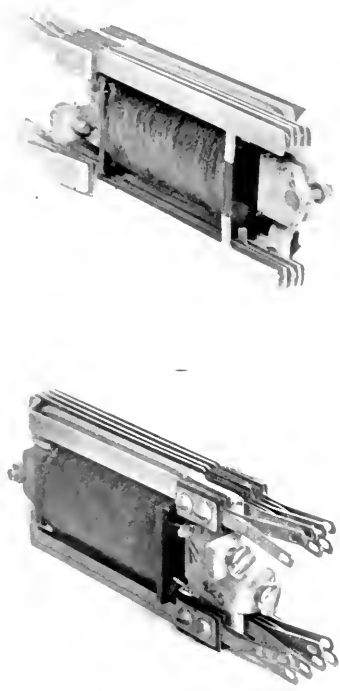


Fig. 13—"E" type relay

in the front of the relay, the cover and the cover cap. By removing the cover cap these screws become accessible and the replacement of the cap does not influence the magnetic conditions or disturb established adjustments.

A General Utility Relay

The success of the punched-type line, cutoff, and supervisory relays suggested the use of this type for a general utility relay which would carry a load of either one pair or several pairs of springs and permit an almost unlimited number of contact spring combinations to be made. This was accomplished by increasing the cross section of the core and armature of the line relay as the increase in iron cross section provided maximum flux with large magnetizing forces. This relay is shown in Fig. 13 and is now manufactured in large quantities with about 3,000 varieties of windings and spring arrangements. About twenty million such relays are already in service and the number is increasing constantly. Had it not been for the development of this punched-type relay, it would have been necessary to greatly increase the manufacturing facilities over those now provided because of the magnitude of the manufacturing operation on the old basis.

Certain Relay Groups

Having outlined the development of the most commonly known relays and given the reasons responsible for major design changes, it will be interesting to consider uses of simple relays in the full mechanical system. In this system the removal of a substation switchhook causes a line relay in the central office to operate and associate a line finder with the calling line, after which a cutoff relay removes the line relay from the circuit as is done in manual practice. A sender is associated with the calling line and the circuit is completed through the substation set dial and a relay in the sender, known as the pulse relay, because it reproduces the dial pulses.

A schematic for illustrating the principle of this circuit is shown in Fig. 14. Referring to this figure, it will be seen that the operation of the pulse relay provides a ground for a slow release relay which in turn extends the circuit of the stepping switch to the back contact of the pulse relay. Suppose that the digit 0 is dialed. Then the resulting current interruptions consist, as shown in Fig. 11, of ten break periods and ten make periods, the final make period being permanent and the remaining nine consisting of approximately one-third of the total time of a single pulse. The first break of the dial opens the circuit through the pulse relay, which releases and opens the circuit of the slow-release relay, but the latter remains operated throughout the break period. The pulse relay when released, provides a ground from its back contact, for the magnet of the stepping switch, through the make contact of the slow release relay. The

stepping switch magnet operates the switch armature and holds it in a position to advance the switch a single step when the magnet is released. When the dial contacts close the circuit again the pulse relay re-operates, releasing the stepping switch, which advances one step, and reestablishing the circuit for the slow-release relay. This cycle is repeated for each break and make pulse period in order to advance the stepping switch over the number of terminals corresponding to the digit dialed.

The adjustment of substation dials is such that pulses are sent at a rate of speed of not less than eight, or more than twelve pulses per second. The break period of individual pulses may vary from .045

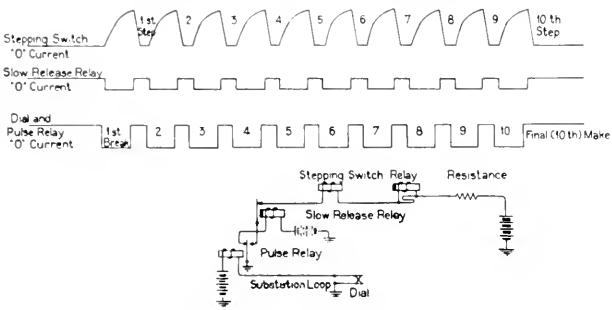


Fig. 14—Curve showing pulsing impulses

to .100 second and the make period may vary from .025 to .050 second. The magnet of the stepping switch must, therefore, complete the movement of the armature in a minimum of .045 second and the switch must advance a single step in a minimum of .025 second. In addition, the slow-release relay must remain operated for a maximum of .100 second; for if it releases during the break-pulse period, the circuit to the stepping switch will be opened. These time values assume that the pulse relay accurately reproduces the dial pulses and it is evident that to accomplish this, its time of operation and release must be independent of the battery potential, between the voltage limits prescribed for the battery, and must also be independent of the differences between the electrical constants of different lengths of substation loops. These are difficult requirements and a punched-type general utility relay, shown in Fig. 13, was used for the purpose as it appeared to be the most suitable avail-

able relay. Its time constant, however, is influenced by the electrical constants of the loop with which it is associated; so that the length of the loop effects the speed at which the armature operates and releases and thus causes the relay to introduce some pulse distortion.

An Accurately Adjustable Flat Type Relay

In order to decrease this distortion a new punched-type relay was designed which reproduces dial pulses with much greater accuracy. It will be seen from the picture of this relay shown in Fig. 15 that the armature is light, that the air gaps can be adjusted closely and with great precision, and that the reduction in the inertia of the armature was obtained by changing the position of the supporting

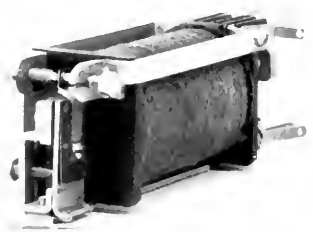


Fig. 15—"L" type pulsing relay

reed hinge. The core of this relay is of small cross section, so that a condition of magnetic saturation is obtained with small current values. With maximum flux on long loops, the increase in current as the length of the substation loop decreases produces very little change in the total flux. Also, changes in the armature air gap as the armature approaches the core do not reduce the reluctance of the magnetic circuit appreciably, so that the armature operates and releases with little time variation irrespective of changes in the electrical constants of the loop.

The slow-release relay, in Fig. 11, is a round-core relay with a reed hinge armature, similar in general construction to the cutoff relay previously described in connection with the early manual system. It is provided with a copper sleeve on the core which acts as a short circuited secondary transformer winding of very low resistance.

Relays in Fundamental Selecting Circuit

For another interesting example of the importance of relay operation in machine switching circuits assume that it is desired to select the fourth terminal in a particular group of a final selector bank as this terminal represents a subscriber's line which has been called from another station. A schematic illustrating the principle of the fundamental circuit for selecting this terminal is shown in Fig. 16. The calling subscriber, by dialing the number of the called station, has

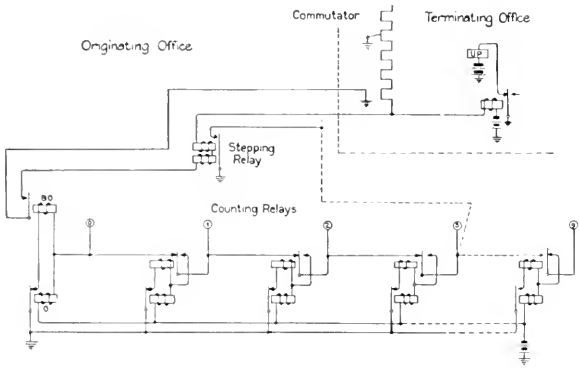


Fig. 16—Schematic of selecting circuit

established the circuit condition shown in this figure through the medium of the sender, so that both the line relay on the final frame in the called office and the stepping relay in the calling office are operated. The circuit is also closed through the up-drive magnet of the final frame, and the selector multiple brush is advancing toward the bank terminal to be chosen. As the selector is driven upward, the commutator brush making contact with the first commutator segment, of the particular group desired, places ground on the interoffice trunk in the called office which shunts down the stepping relay in the calling office. This releases the stepping relay, which had established a circuit when operated through the lower relay of the fourth pair of counting relays and had shunted the upper relay of the pair so it would not operate. The release of the stepping relay removed this shunt and permitted the upper relay to operate, locking both relays through the make contact of the lower relay and trans-

ferring the start lead to the lower relay of the third group which operates when the start lead is again grounded through the make contact of the stepping relay. This cycle is repeated for each segment which the commutator brush passes over until the upper relay of the fourth or zero group of counting relays operates and opens the fundamental selecting circuit, thus allowing the line relay in the final frame to release when the commutator brush again removes the shunt. The line relay, on releasing, opens the up-drive circuit and the selector stops with the multiple brush resting on the particular terminal desired.

There are three different types of relays in this circuit. The line relay on the final frame is the general utility punched-type relay of Fig. 13 with the contact spring assembly and mechanical adjustments required by the specific circuit condition. It is evident that this relay must release quickly enough to enable the up-drive clutch magnet to release before the selector is driven beyond the desired terminal or a false bank terminal selection will be made. An examination of some of the factors influencing the release time of the line relay will therefore be of interest.

When the commutator brush made contact with the commutator segment both ends of the inter-office trunk were grounded but before the brush left this segment the condenser charge on the trunk leads was dissipated and the distant end of the trunk was opened by the operation of the upper counting relay of the zero group. On leaving the fourth commutator segment the brush opened the circuit of the line relay which could not release instantaneously because of its own time constant, the transient current through its windings for charging the trunk capacity, and the leak current in its windings resulting from trunk leakage.

The time constant is determined by the electrical and magnetic constants of the relay and for a given winding is inherent to its structure. If the time constant is such that adjustments, for armature air gap, spring tension and contact separation, cannot be made which will enable a relay to meet all the circuit requirements, a different type of relay structure having a more favorable time constant must be used.

The magnitude of the charging current for the trunk is determined by the trunk capacity and is in direct proportion to the length of the trunk which is limited to 12 miles corresponding to a maximum capacity of about 0.84 mf. The limiting open circuit resistance of the trunk is 30,000 ohms and the standard of maintenance is such that the insulation resistance is not allowed to drop below this value.

In addition the maximum resistance of the trunk is 1300 ohms. The line relay must therefore be adjusted to operate over this resistance and in series with the stepping relay when the battery potential is a minimum of 44 volts. It must also be adjusted to release quickly enough to insure the positive selection of a particular terminal when the battery potential is a maximum of 52 volts and both the trunk capacity and trunk leakage are maximum. These are very severe requirements to be met by a relay which is produced commercially in large quantities at a small cost; and more severe conditions such as would result, for example, from increasing the length of the trunk could not be imposed on this particular relay unless the iron structure were made from some new material having more favorable magnetic constants.

The requirements for the stepping relay, however, are more severe than those for the line relay, for the stepping relay must continually operate and release as the commutator brush alternately grounds and frees the trunk in the distant office. Also the insulation resistance and capacity of the trunk exert a somewhat different influence on the functioning of the stepping relay than on the functioning of the line relay. The trunk leakage current resulting from low insulation resistance interferes with the operation of the stepping relay, instead of its release, so it must be adjusted to operate on a minimum battery potential of 44 volts and a minimum trunk insulation resistance of 30,000 ohms. The trunk capacity interferes more seriously with the release of the stepping relay than with the release of the line relay. When the ground is removed from the latter the trunk is at zero potential and the charging current through the relay windings is maintained for a very brief period of time but when the incoming end of the trunk is grounded to release the stepping relay in the distant office, the trunk capacity is fully charged and the discharging current is sustained for a much longer time interval.

The Stepping Relay

The time constant of the line relay is such that it cannot be given adjustments which will enable it to meet the more severe requirements of the stepping relay, and consequently an entirely different type of structure, as shown in Fig. 17, is used for a stepping relay. This design is of particular interest because it is not used for any other purpose and is the only relay of its type in the telephone plant. Many attempts have been made to replace it with some sort of punched type structure that is more adaptable to the established manufactur-

ing methods but they have been ineffectual as yet, for the equivalent combination of sensitivity and reliability and a delicate means of adjustment is difficult to attain. In order to satisfy the severe circuit conditions the stepping relay is adjusted to operate on 10 mil-amperes and not to operate on 9 mil-amperes, a difference of only 10 per cent. in the operate and non-operate adjustments.

The stepping relay must reproduce the pulsations of current originated by the commutator brush with sufficient accuracy to insure

Fig. 17—Stepping relay

the positive operation of the counting relays, for any failure of the latter will result in false selection. The stepping relay must therefore maintain a circuit through its make contact for a sufficient time to enable the lower relay of any counting pair to operate and must open the circuit through the same contact long enough to permit the upper relay of the pair to lock up in series with the lower relay. Since the stepping relay does not always reproduce the commutator pulses perfectly and since any pulse distortion must necessarily reduce the operating time margin for one of the relays of a counting pair, it is evident that rapid operation and reliability of operation are essential characteristics for the counting relays. A punched type relay similar to the line relay cannot operate with sufficient speed. The stepping relay would qualify for speed, but a complete set would require considerable space and would be inconvenient to mount.

The Counting Relay

The relay designed for a counting relay is shown in Fig. 18 and has the qualities of speed and reliability that are required. It is equipped with a light armature, on a pivot suspension, that operates

Fig. 18—Counting relay

through a small air gap. The contacts are mounted on rigid springs that cannot be adjusted readily, but which maintain a given adjustment, without change, for a long time. This relay, like the stepping

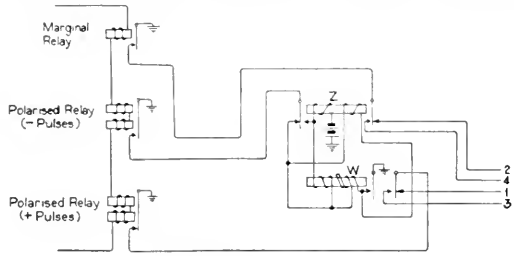


Fig. 19—Call indicator circuit

relay, is unique, in that, it is not used for any other purpose in the telephone system, and in addition all attempts to design a punched type relay that is a satisfactory substitute have, so far, been unsuccessful.

Certain Marginal and Polarized Relays

Another interesting and unusual use of relays is the arrangement at the terminating end of a call indicator trunk from a full mechanical to a manual office. This arrangement consists of three relays in series in the manual office, as shown in Fig. 19. One of them is a

marginal relay adjusted to operate on any current greater than a particular value. The other two are polarized relays, one being adjusted to respond to negative pulses only, while the other responds only to positive pulses.

Each digit of any number transmitted over the trunk to the manual office consists of four pulses. The second and fourth of these pulses are always negative, but either or both of them may be a light or heavy negative. The first and third pulses may either be positive or zero, a zero pulse representing a no current interval. This combination of pulses is shown in the following table:

1	2	3	4
+		+	-
0	-	0	-

As each pulse interval may consist of either of two kinds of pulses, there are sixteen combinations which can be transmitted but six of them are not used, as only ten are required.

The marginal relay is adjusted to operate on heavy pulses only and as all the positive pulses are light, it does not respond to any positive pulses or the light negative pulses. The negative polarized relay responds to both light and heavy negative pulses and the positive polarized relay responds to all positive pulses. During a zero pulse period all of the relays remain unoperated. For the second and fourth pulse periods the negative polarized relay will be operated and the marginal relay may or may not be operated. During the first and third pulse periods the positive polarized relay may be operated or all the relays may remain unoperated. From the operation of these relays an arrangement of register relays is set up which lights before the manual operator the lamps corresponding to the digits transmitted. The marginal relay used is a counting relay of the type shown in Fig. 18, as this relay has the qualities of sensitiveness, stability and permanence of adjustment that are essential for satisfactory operation. The other two relays are very sensitive polarized relays with micrometer adjustment screws and are representative of the best standards of design for relays of their type. This type of relay is shown in Fig. 20.

The Sequence Switch

Most of the relays previously described were designed to meet specific requirements of unusual severity which limited the design to individual structures having their armatures in close association

with the contact carrying springs. Many of the switching operations required for relaying a circuit from point to point through an office can be performed under conditions allowing greater latitude in relay design which has led to the development of several interesting and unusual forms of multi-contact relays in which the armatures

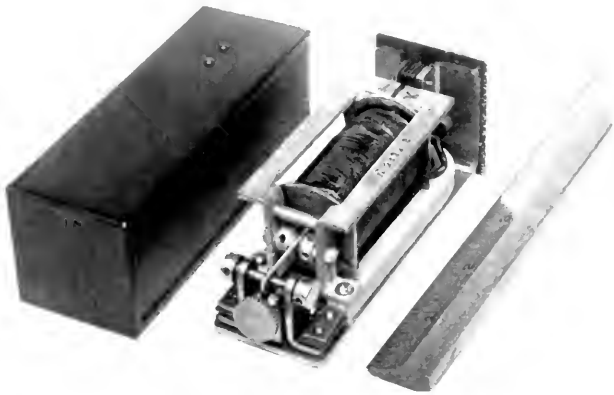


Fig. 20—Call indicator polar relay

indirectly control groups of contact carrying springs. In the development of the machine switching system the work of establishing circuits performed by human relays was transferred to mechanical relays and it soon became evident that the number of individual relay structures of the conventional type required for such a substitution would be so great and the circuit arrangements would be so complicated that the cost would be prohibitive.

The 24 cam sequence switch shown in Fig. 21 is an interesting example of the remote contact control multi-contact relay that not only performs the functions of a multitude of individual relays but actually replaces entire circuits which would require large numbers of relays to control the particular relays that transferred the circuit from point to point. The relay sequence switch shown in the figure is assembled with a shaft that may be rotated into any one of 18 positions which are stamped on an index wheel and are indicated by the position of the wheel with reference to a pointer fixed to the frame

of the switch. Each of the circuit switching cams is associated with four brushes and it is possible to so arrange the contact carrying segments on these cams that 6624 different circuit combinations can be established by advancing the switch successively into each of the 18 positions.

The switch is propelled by a driving disc mounted on a power driven shaft that revolves constantly at a speed of 36 r.p.m. The driven disc on the switch in association with the driving disc constitutes a

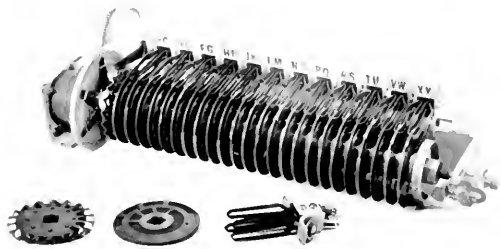


Fig. 21—24 cam sequence switch

friction clutch under the control of an electro-magnet which deflects the driven disc to bring it into relation with the driving disc when it is desired to advance the switch. The electro-magnet corresponds to the winding of an individual relay structure and the driven disc is the armature, the combination of the winding and armature simply serving as a means for controlling the contact relations of a multiplicity of springs.

The Power Driven Selector

The power driven selector shown in Fig. 22 is another example of an entirely different form of multi-contact relay for transferring the three contacts of any one of 500 circuits to the contact springs of a brush that will relay that circuit to any desired point. These 500 circuits are assembled in five groups of 100 each in five banks that are mounted on a frame as shown in the figure. Five brushes, one for each bank, are assembled on a vertical rod in such relation to the banks that the mechanical tripping or release, of any brush brings its springs in contact with the terminals of the bank with which it is associated. The corresponding springs of each of the five brushes

are connected in multiple so that in relaying a circuit it is necessary to trip only that brush which is presented to the bank in which the terminals appear. Bringing the brush springs in contact with a par-

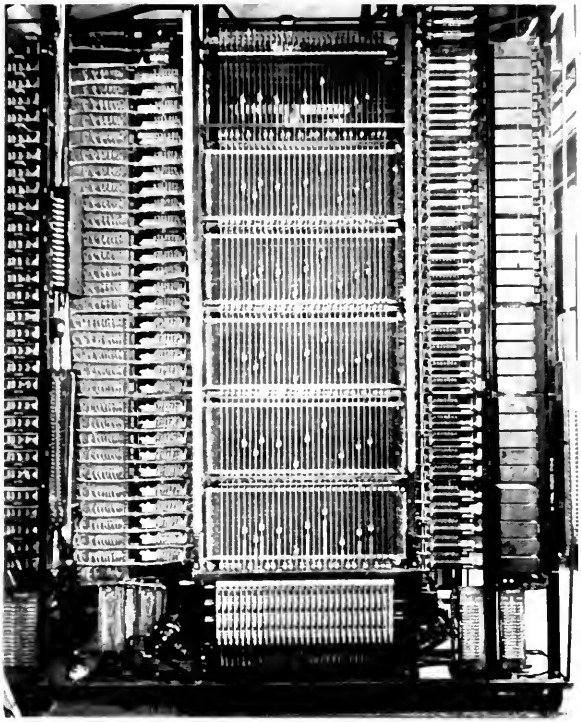


Fig. 22—Power driven selector

ticular group of terminals is referred to as a process of selection and is accomplished by driving the brush rod upward until the desired terminals are reached. When the circuit arrangement is no longer desired the brush rod is driven downward to a normal position where the tripped brush is also restored mechanically to its original condition.

The power for elevating and restoring the brush rod is provided by continuously revolving motor driven steel rolls covered with cork and mounted at the base of the frame. The driving of the brush rod, the tripping of the desired brush, the stopping of the rod and its restoration to normal are all controlled by a series of electro-magnets assembled in a single structure called a clutch which is also mounted at the base of the frame directly in front of the rolls. When a brush rod is driven either up or down, a clutch armature establishes a friction contact between a flat strip of phosphor bronze fastened to the lower end of the brush rod and the cork on the revolving rolls. This clutch is comparable to an individual relay structure with a multiplicity of windings and armatures that are so related that each armature will operate only when its associated winding is energized. The clutch thus does the work of either an exceedingly intricate individual relay or a whole group of less complicated relays. The clutch windings are in effect, relay windings that control the positions of remote contact springs through the operation of armatures which associate or disassociate electro-magnetic and mechanical energy as is desired.

The Step-by-Step Selector

Another type of multi-contact relay in general use that differs in form from both the sequence switch and the power driven selector is the step-by-step selector shown in Fig. 23. It consists of six semi-circular contact levels assembled in a bank and an electro-magnet which drives a set of six, double ended, rotary brushes over the terminal arc by means of a driving pawl and ratchet wheel. Each time the magnet is energized and released the driving pawl engages the next tooth on the ratchet wheel which rotates to advance the brushes a single step so that they make contact with the next set of terminals. In 11 successive steps the six brushes move through a complete revolution but as they are double-ended all the possible circuit combinations are set up in the first 22 steps and are then repeated.

In this selector the winding of the electro-magnet corresponds to the winding of an individual relay. The armature in operating elongates a spring that is shown in Fig. 23 and the energy stored in this spring restores the armature to normal and advances the six contact making brushes to the next set of contact terminals. Thus the relay winding and armature control the position of the contact springs through the agency of a flexible mechanical link. The relay winding may be alternately energized and released by current interruptions from an outside source or the armature may be arranged to

interrupt the circuit through the winding by opening a pair of contacts in the operated position to advance the selector by self interruptions.

Relays in Toll Circuits

Supervision on all of the longer toll circuits and on most of the shorter ones is provided on what is known as a ringdown basis. This

Fig. 23—Step-by-step rotary switch

usually involves a ringup relay at each end of the line, which operates in response to 20-cycle signaling impulses. These impulses may be transmitted over the line from one office to the other or they may originate in the same office as the relay and be impressed on the line by the operation of a so-called composite ringer in response to signals of a different frequency. The ringup or drop relay provides the signal in the toll switchboard. It is usually removed from the circuit when the line is taken up by the operator and the supervision is then transferred to the toll cord.

The toll cord supervisory circuit is shown in Fig. 24 and illustrates a typical condition which has imposed particular requirements on the relays involved. The signal receiving relay *A* may be bridged

directly across the line conductors, in which case its winding must be of such high impedance that it does not materially affect the efficiency of the talking circuit. Experience has indicated that with the windings commonly used on relays there is some chance of sufficient short circuited turns to materially reduce the inductance of the winding. This may occur in a relay which would otherwise give satisfactory operation, the short circuited turns merely reducing

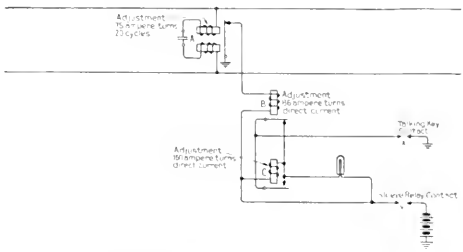


Fig. 24—Toll cord supervisory circuit

slightly the low frequency or direct-current efficiency. For this reason, the relay winding has been divided into two parts, on separate cores, either one of which has sufficient inductance to safeguard the telephone transmission.

The incoming 20-cycle signaling current may be of small value and the portion through the relatively high impedance of the relay will be still smaller so that this relay must be extremely sensitive. The relay has small moving parts and a comparatively light spring tension. These factors contribute to sensitive operation but also permit the opening of the contact on impulses other than those intended for signaling. Such impulses are usually of short duration and the other relays of the circuit have been designed to limit their effect to prevent false signals.

Both relays "*B*" and "*C*" are of the same type, designed to operate with a slight time lag so that other things being equal they would be expected to operate at the same time when the circuit is closed at *x* and *y*. Relay "*B*," however, receives, under the worst condition, 150 per cent. of its rated operating current, while relay "*C*" receives 105 per cent. This will tend to make relay "B" quicker in operation than relay "*C*," so that when the battery and ground are connected to the circuit, relay "*B*" will operate first and open the winding of relay "*C*." This is therefore the normal condition of the

circuit and is further insured by the fact that the opening of the winding of "C" occurs at a back contact of relay "B" while the locking of "C" occurs only after the relay has pulled up to close its front contact.

The sequence of operation and release resulting from this series of relay operations affords protection against false signals since relay "A" must operate continuously until "B" has released and "C" has operated before the lamp circuit is closed. Relay "B," in addition to being slow in operation, is also slow to release, so that the time interval thus introduced tends to bridge over any transient impulses that may tend to operate the signal.

The slow operation of relays is secured by means of a copper sleeve over the relay core. Slow operation results from the transient condition existing during the time between the application of voltage to the relay winding and the building up of the magnetic field to a steady state. Slow release results from the transient condition existing during the time between the removal of the voltage from the relay winding and the decay of the magnetic field until the magneto-motive force falls below the armature restoring force. These conditions are more easily seen when the relay winding is considered as the primary of a transformer and the copper sleeve as a short-circuited secondary winding consisting of a single turn having a very low resistance. The operating current, before it reaches its steady value, may be considered as an alternating current of one-quarter of a cycle, starting from zero and building up to a maximum value. Slow operation of the armature results from opposing the building up of the flux in the core. Slow release is due to retarding the decay of the flux in the core. The speed at which the armature operates or releases is not changed but in the first case the application of the magneto-motive force required to move the armature is delayed, and in the second case the removal of the magneto-motive force holding the armature in the operated position is also delayed. When a voltage is first applied to the terminals of the winding, the current tends to build up and establish the magnetic flux at its maximum value in the relay core. The instant the flux threads the copper sleeve, a voltage is induced in the latter, causing a current to flow in it. This current in the copper sleeve sets up a flux in the same magnetic path which opposes the flux building up from the current in the relay winding. Due to leakage, the winding flux is greater than the opposing flux set up by the sleeve and the resultant flux continues to build up until it reaches a maximum value. This opposition of the winding flux and the flux produced by the induced current in the copper sleeve

increases the time for the building up of the magnetic force necessary to move the armature from the normal position. It also increases the time for such a reduction in the magnetic force as will permit the armature to release.

The slow release feature is further secured by omitting the stop pins which are usually provided between the armature and the pole

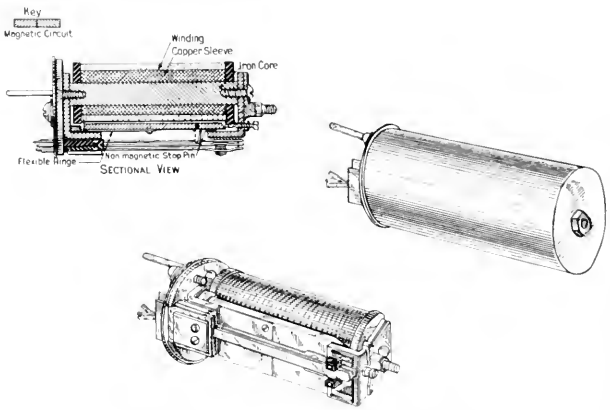


Fig. 25—Sectional view of No. 162 type relay, a slow operating relay

piece. This tends to delay the decline in flux through the magnetic circuit when the current is interrupted. Fig. 25 illustrates diagrammatically the structure of these relays.

Relays of the Composite Ringer

A somewhat similar use of relays is to be seen in the composite ringer circuit mentioned above. The relay circuit of such a ringer is shown in Fig. 26, in simplified form. This circuit is designed to receive 20-cycle signals from the switchboard and transmit out on the line signals of a higher frequency and to receive the higher frequency impulses and in turn transmit 20 cycles to the switchboard. In this case, the 20-cycle relay "*A*" does not meet the requirement for high impedance since protection to the telephone circuit is afforded by coil "*C*." A single core is therefore satisfactory and a positive make-contact relay is used. In this case, the chief requirement is that relay "*B*" should be slow in operating.

The chain of relays operating from the high frequency signals consists of relays *D*, *E* and *F*. Relay "*D*" must be a very sensitive structure in this case and a polarized relay with a vibrating contact has commonly been used. The circuit requirements are such that the energy available for the operation of this relay is seldom more than a few hundred microwatts and may be much less. The cir-

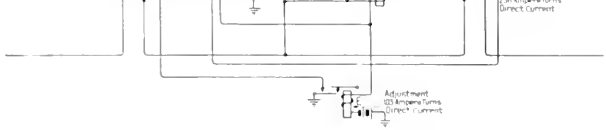


Fig. 26—Composite ringer circuit

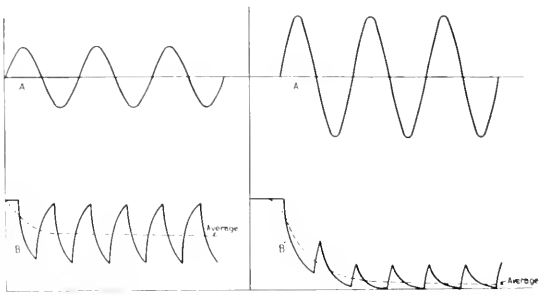
cuits are being designed on the basis of giving reliable operation on 20 microwatts. The operation of relay "*D*" releases relay "*E*" which in turn operates relay *F*.

Where such a circuit depends on the operation of a vibrating contact relay, the current through this contact is of vital importance. Whenever the contact is closed, current tends to flow through the winding of relay "*E*." Fig. 27 illustrates the effect of very weak signaling currents and of currents sufficient to give proper operation. The current values through the vibrating relay winding and through the winding of the secondary relay are shown for two different typical conditions. Also, the average or effective value in winding "*E*" is shown.

A circuit feature which has recently been introduced to increase the sensitivity of relay "*D*" and to improve the operation of the secondary relay consists in the introduction of a condenser and the operation of the vibrating contact as a normally open contact. The closing of the contact charges a condenser which tends to operate the secondary relay by its discharge as soon as the contact opens. By this combination, the effect noted in Fig. 27 is eliminated and positive operation of the secondary relay is secured as soon as the armature vibrates sufficiently to make contact. The local circuit embodying this feature is shown in Fig. 28. Referring to this figure and to Fig. 26, relays "*D*" and "*E*" represent the alternating current

relay and the secondary relay in each case. In the one case, however, relay E' operating when relay "D" operates gives positive release of relay "E" instead of introducing an uncertain resistance in its circuit.

This circuit embodies several features which are not common in relay systems. The operation of relay E' is dependent on the values



CURRENT IN A.C.SIGNALING RELAY
AA Current in Winding
BB " through Contact

Fig. 27.—Curve showing signal impulses in a.c. signaling relay
AA', current in winding
BB', current through contact

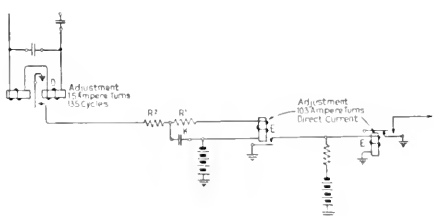


Fig. 28.—Circuit for a make contact 135 cycle relay

selected for resistances R' and R^2 and for condenser K. These values must be such that the current in the relay winding is maintained during the opening of the contact of relay "D" by means of the discharge current from the condenser. On the other hand K and R^2 must be proportioned to limit the arcing of the contact of "D" at

the frequency of the signaling impulses. The method of releasing relay "*E*" by short-circuiting its winding has advantages over opening the circuit for the purpose under consideration. The arcing at the contact of relay *E* is less severe than would be the case if an inductive circuit were broken.

An added feature which has been incorporated in the mechanical design of relay "*D*" and which has an important bearing on its performance electrically, is an adjustment limiting the armature travel. This limitation of movement prevents a wide deflection when the relay receives excessive current. Such deflection would tend to set the armature into vibration and would result in a sufficient number of impulses to operate relay *E* and cause false signals.

The Vacuum Tube

The vacuum tube is used for the relaying of energy in a number of ways. It may be connected in circuit to amplify the received impulses in which case it sends out energy from a local source with the same wave shape as that of the received current. In this case the tube serves to relay the impulses with as little distortion as possible. In the case of a tube used as a modulator or a demodulator it is required to combine or separate impulses of different character, the two operating together to preserve the same impulses at the output of the demodulator tube as is received at the input of the modulator. The impulses which are transmitted between the two tubes have an entirely different wave form and may be amplified any number of times by means of amplifier tubes without affecting the action of the modulator and demodulator.

The vacuum tube may also be used as a rectifier to convert alternating current to direct or pulsating current or it may be used as an oscillator to produce alternating current from a local source of direct current. In all of these applications of vacuum tubes, the tubes serve as relays to introduce a fresh supply of energy or a desired wave form or a combination of the two to serve their purposes in the communication system.

Relays for Telegraph Circuits

The use of relays for telegraph circuits presents an entirely different set of problems than those usually encountered in the consideration of telephone circuits. Most telegraph relays are used for repeating signals from one circuit to another rather than for switching local circuits. While some marginal operating conditions are

imposed on telephone relays there is not the wide range of operating conditions to be met under which most telegraph relays are required to operate. The numbers involved are usually much less so that economies in production play a somewhat less important role and the cost is not quite such an important item. Similarly the methods of assembly and mounting afford a somewhat wider latitude than can be permitted where many thousands of relays must be mounted in comparatively small space.

Because of the exacting requirements imposed on telegraph relays and to insure continuity of service as far as possible, they are usually made interchangeable to a much greater degree than telephone relays. They may be connected by means of screws instead of soldered connections or they may be inserted in the circuit by means of spring clips in a connecting block.

In a telegraph system speed of operation and reliability are the most important requirements and are very large factors in determining the mechanical design of the relays. The relay must operate quickly and accurately so as to cause as little distortion as possible to the signals. In addition it must be extremely rugged and maintain its adjustment well throughout long continued operation. A very ordinary day's work for a telegraph relay requires the reliable operation of its contacts several hundred thousand times and it may be called upon to open and close its contacts a million times a day. Where a telephone relay might hesitate and still pull up and perform its function properly or might make uncertain contact at first, such behavior on the part of a telegraph relay would result in false impulses and would quickly call for a readjustment or a change of relays.

With the exception of some of the alternating current signaling relays in telephone circuits the energy available in telegraph relays is usually less than that available for telephone relays. The more sensitive relays are called upon to operate from line current which has been attenuated by leakage or by parallel paths and which may have been limited at the distant station. Systems operating over open wire lines are usually restricted to about .075 ampere at the sending end and in cable the normal current is about .005 ampere. This difference is not as great in actual operation as would appear since the open wire system operates on a ground to ground basis and the cable on a metallic basis. In operating from ground at one station to ground at another differences in ground potential and leakages occur which require a greater margin than is necessary with the metallic system.

If satisfactory telegraph service is to be rendered, particularly on long circuits involving a number of relaying points, it is essential that the telegraph relays employed have such operating characteristics that they introduce as little distortion to the signals as possible. It has been found that the polarized type of relay fulfills this condition to a greater extent than the neutral type of relay which is used in local circuits and in some telegraph circuits where extreme accuracy is not required. The polar relay permits arrangements of circuits which minimize the effect of poor wave shape and line leakage. It also is more easily adapted to variations in current strength and may be adjusted to give more accurate repetition of the signals under all conditions.

A number of important developments in telegraph relays have led up to the relay shown in Fig. 29. This relay gives reliable operation with 1-ampere turns in the winding and by careful adjustment may be made to operate on a small fraction of that.

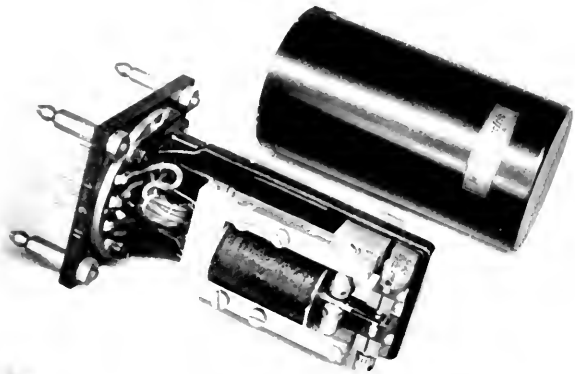


Fig. 29—Photograph of telegraph type relay

While the telegraph relay may be called upon to operate on very small energy, its contact must be capable of handling much larger quantities. Due to the speed of operation desired and to the dependence on accurate transmittal of each impulse the contacts must operate without chatter or vibration. Great care has been taken in the design of the relays and the circuits to protect the contacts and

insure good operation. Chatter can be largely eliminated by careful mechanical design and the effect of the arc set up when the contact is called upon to break the current in a circuit carrying several watts can be minimized somewhat by means of, so-called, spark killers. These consist of condensers and resistances so proportioned as to absorb the force of the arc in the charging of the condenser when the contact opens. They can be utilized still further in modifying the shape of the transmitted wave by the charging of the condenser when the contact is opened.

Telegraph relays and their applications have been referred to in this paper only in the most general terms because of the variety of their forms and uses. It is planned to cover this as well as other subjects pertaining to relays in a series of later papers.

Some Applications of Statistical Methods to the Analysis of Physical and Engineering Data

By W. A. SHEWHART

Synopsis—Whenever we measure any physical quantity we customarily obtain as many different values as there are observations. From a consideration of these measurements we must determine the *most probable value;* we must find out *how much* an observation may be expected to vary from this most probable value; and we must learn as much as possible of the *reasons why* it varies in the particular way that it does. In other words, the real value of physical measurements lies in the fact that from them it is possible to determine something of the nature of the results to be expected if the series of observations is repeated. The best use can be made of the data if we can find from them the most probable frequency or occurrence of any observed magnitude of the physical quantity or, in other words, the most probable law of distribution.

It is customary practice in connection with physical and engineering measurements to assume that the arithmetic mean of the observations is the most probable value and that the frequency of occurrence of deviations from this mean is in accord with the Gaussian or normal law of error which lies at the foundation of the theory of errors. In most of those cases where the observed distributions of deviations have been compared with the theoretical ones based on the assumption of this law, it has been found highly improbable that the groups of observations could have arisen from systems of causes consistent with the normal law. Furthermore, even upon an a priori basis the normal law is a very limited case of a more generalized one.

Therefore, in order to find the probability of the occurrence of a deviation of a given magnitude, it is necessary in most instances to find the theoretical distribution which is more probable than that given by the normal law. The present paper deals with the application of elementary statistical methods for finding this *best* frequency distribution of the deviations. In other words, the present paper points out some of the limitations of the theory of errors, based upon the normal law, in the analysis of physical and engineering data; it suggests methods for overcoming these difficulties by basing the analysis upon a more generalized law of error; it reviews the methods for finding the best theoretical distribution and closes with a discussion of the magnitude of the advantages to be gained by either the physicist or the engineer from an application of the methods reviewed herein.

Introduction

We ordinarily think of the physical and engineering sciences as being exact. In a majority of physical measurements this is practically true. It is possible to control the causes of variation so that the resultant deviations of the observations from their arithmetic mean are small in comparison therewith. In the theory of measurements we often refer to the "*true value*" of a physical quantity; observed deviations are considered to be produced by errors existing in the method of making the measurements.

With the introduction of the molecular theory and the theory of quanta, it has been necessary to modify some of our older conceptions. Thus, more and more we are led to consider the problem of measuring any physical quantity as that of establishing its most probable value. We are led to conceive of the physico-chemical laws as a statistical determinism to which "the law of great numbers" [1] imparts the appearance of infinite precision. In order to obtain a more comprehensive understanding of the laws of nature it is becoming more necessary to consider not only the average value but also the variations of the separate observations therefrom. As a result, the application of the theory of probabilities is receiving renewed impetus in the fields of physics and physical chemistry.

Statistical Nature of Certain Physical Problems. As typical of the newer type of physical problem, we may refer to certain data given by Prof. Rutherford and H. Geiger.[2] In this experiment the number of alpha particles striking, within a given interval, a screen subtending a fixed solid angle was counted. Two thousand six hundred and eight observations of this number were made. The first column of Table I records the number of alpha particles striking this screen within a given interval. The second column gives the frequency of occurrence corresponding to the different numbers in the first column.

TABLE I

No. of Alpha Particles	Observed Frequency of Occurrence
0	57
1	203
2	383
3	525
4	532
5	408
6	273
7	139
8	45
9	27
10	10
11	4
12	0
13	1
14	1

It is obviously impossible from the nature of the experiment to attribute the variations in the observed numbers to errors of observation. Instead, the variations are inherent in the statistical nature of the phenomenon under observation.

[1] Each class of event eventually occurs in an apparently definite proportion of cases. The constancy of this proportion increases as the number of cases increases.

[2] *Philosophical Magazine*, October, 1910.

The questions which must be answered from a consideration of these data are typical. For example, we are interested to know how a second series of observations may be expected to differ if the same experiment were repeated. The largest observed frequency corresponds to four alpha particles, although what assurance is there that this is the most probable number? What is the probability that any given number of alpha particles will strike the screen in the same interval of time? Or again, what is the maximum number of alpha particles that may be expected to strike the screen? All of these questions naturally can be answered providing we can determine the most probable frequency distribution.

Statistical Nature of Certain Telephone Problems. The characteristics of some telephone equipment cannot be controlled within narrow limits much better than the distribution of alpha particles could be controlled in the above experiment. We shall confine our attention primarily to a single piece of equipment. The carbon microphone. For many reasons it is necessary to attain a picture of the way in which a microphone operates. It is necessary to find out why carbon is the best known microphonic material. In order to do this we must measure certain physical and chemical characteristics of the carbon and compare these with its microphonic properties when used under commercial conditions. In the second place it becomes necessary to establish methods for inspecting manufactured product in order to take account of any inherent variability, and yet not to overlook any evidence of a "trend" in the process of manufacture toward the production of a poor quality of apparatus. In the third place it so happens that the commercial measure of the degree of control exhibited in the manufacture of the apparatus must be interpreted ultimately in terms of sensation measures given by the human ear. That is, the first phase of the problem is purely physical; the second is one of manufacturing control and inspection and the third involves the study of a variable quantity by means of a method of measurement which in itself introduces large variations in the observations.

In one of the most widely used types of microphones there are approximately 50,000 granules of carbon per instrument. Each of these granules is irregular in contour, porous and of approximately the size of the head of a pin. If such a group of granules is placed in a cylindrical lavite chamber about $\frac{1}{2}$-inch in diameter and closed at either end with gold-plated electrodes; if this chamber is then placed on a suspension free from all building vibrations and carefully insulated from sound disturbances; if automatically controlled

mechanical means are provided for rolling this chamber at any desired speed; if all of the air and sorbed gases are removed from the carbon chamber and pure nitrogen is substituted; if the mean temperature is kept constant within 2° C; and if means are provided for measuring the resistance of the granules when at rest by observing the voltage across the two electrodes while current is allowed to flow for a period less than 1/200 of a second, it is found that the resistance (for most samples of carbon) may be determined within a fraction of one per

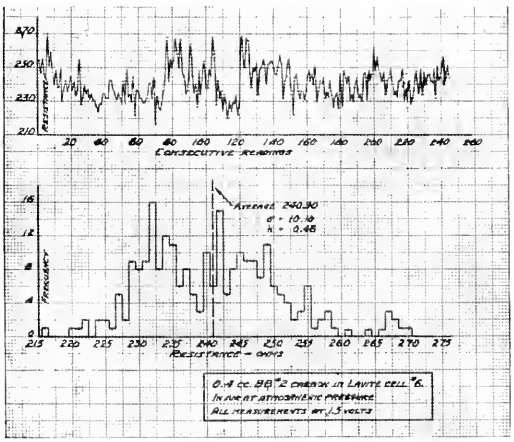


Fig. 1

cent. If, however, the button is rotated (even as slowly as possible) and then brought to rest, the resistance may differ several ohms from its first value. If a large number of observations are made after this fashion, we may expect to find for certain samples of carbon a set of values such as given in Fig. 1. The 270 observations of resistance reproduced in this figure were made on a sample of carbon at $1\frac{1}{2}$ volts under conditions quite similar to those outlined above. The observed variation is from approximately 215 to 270 ohms. The upper curve is that of the resistance vs. the serial number of the readings. There is no apparent trend in the change of resistance from one reading to another. The lower curve in this figure shows the frequency histogram of the results. Attention is directed to the

wide variation in the observations, and to the fact that the frequency histogram appears to be bimodal.[3] Methods of dealing with such distributions will be considered.

Samples of carbon having different molecular surface structures have different resistances. To put it in a still more practical way, if the manufacturing process is not controlled within very narrow limits, wide variations are produced in the molecular properties of the carbon. The microphonic properties of these carbons are therefore different. One of the problems with which we have been concerned is to determine the relationship existing between the physical and chemical characteristics of the carbon and the resistance of the material when measured under different conditions. We are obviously dealing in this case with problems involving the measurement of physical quantities which cannot be controlled even in the labora-

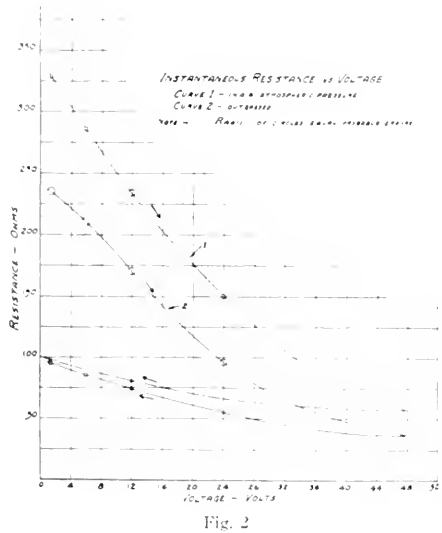


Fig. 2

[3] If curves which touch the axis at $+\infty$ and $-\infty$ have more than one value of the variable for which the derivative of the frequency in respect to the variable is equal to zero the points being other than that for which the frequency is zero these curves are referred to as bimodal, trimodal, etc. The modal value is the most probable one and is of particular interest in unimodal curves.

tory. If we remove the air and measure the resistance at different voltages, we may expect to find changes in the resistance similar to those indicated in Fig. 2. Curves 1 and 2 were taken for increasing voltages. The return curves were taken with decreasing voltage. Removal of the air from this particular sample of carbon produces comparatively large changes in the resistance. The resistance at $1\frac{1}{2}$ volts is several times that at 48 volts. These curves were taken under conditions wherein all of the other factors were controlled. A sufficient number of observations was made in each case in order to establish the probable errors of the points as indicated by the radii of

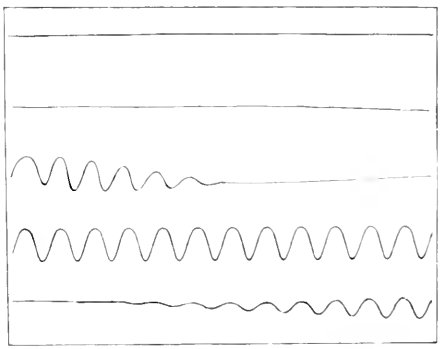


Fig. 3—Possible Types of Breathing of Granular Carbon Microphone.

the circles. If this same experiment were carried on at a different temperature, radically different results would be obtained.

If, instead of allowing the current to flow for a short interval of time, a continuous record is made of the resistance of the carbon while practically constant current flows through the carbon, the resistance will be found to vary. The maximum resistance reached in certain instances may amount to several times the minimum value. In general, this phenomenon is attributed to the effects of gas sorbed on the surface of the material. Transmitters cannot be made of lavite so that the expansions and contractions of the piece parts thereof augment the changes in resistance. This phenomenon, termed "breathing," may be, but seldom is, regular or periodic. An exceptional case of breathing is shown in Fig. 3. This was obtained with a special type of carbon in a commercial structure. The curves

themselves represent the current through the transmitter and, therefore, are inversely proportional to the resistance. All five curves were obtained with the same carbon in the same chamber by varying merely the configuration of the granules by slightly tapping the carbon chamber.

All of these effects can be modified to a large extent by varying the process of manufacture of the granular material. In practice it is necessary to know why slight changes in the manufacturing process cause large variations in the resistance characteristics of the carbon. The same process that improves one microphonic property may prove a detriment to another. It is in the solution of some of these problems that statistical methods have been found to be of great value in the interpretation of the results.

Whereas the physicist ordinarily works in the laboratory under controlled conditions, the engineer must work under commercial conditions where it is often impractical to secure the same degree of control. More than 1,500,000 transmitters are manufactured every year by the Bell System. Causes of variation other than those introduced by the carbon help to control the transmitter. For example, variations may be introduced by the process of assembly, or by differences in the piece parts of the assembled instrument. The measure of the faithfulness and efficiency of reproduction depends fundamentally upon the human ear. Obviously all transmitters cannot be tested. Instead, we must choose a number of instruments and from observations made on these determine whether or not there is any trend in the manufactured product. Naturally we may expect to find certain variations in the results according to the rules of chance. To take the simplest illustration, we may flip a coin 6 times. Even if it is symmetrical we may expect occasionally to find all heads and occasionally all tails, although the most probable combination is that of 3 heads and 3 tails. We must, therefore, determine first of all whether or not the observed variations are consistent with those due to sampling according to the laws of chance. If there is an apparent trend in product, the data should be analyzed in order to determine, if possible, whether it is due to lack of control in the manufacture of carbon or to some other set of causes such as mentioned above. Because of economic reasons we must keep the number of observations at a minimum consistent with a satisfactory control of the product. Here again it has been found that the application of statistical methods is necessary to the solution of the problems involved.

Before considering the problem of the measurement of efficiency and quality of the transmitter, let us consider the schematic diagram

of the telephone system as shown in Fig. 4. Essentially this consists of the transmitter, the line and the receiver. The oldest method of measurement is to compare one transmitter against a standard in the following way. An observer calls first in the standard and then in the test transmitter, while another observer at the receiving end judges the faithfulness of reproduction. The pressure wave striking the transmitter diaphragm varies with the observer and also with the degree of mechanical coupling between the sound source

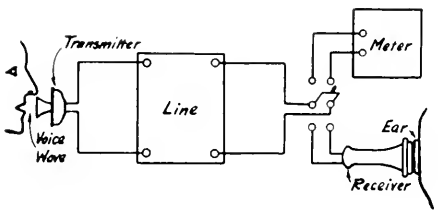


Fig. 4

and the diaphragm of the instrument. The judgment of the observer at the receiving end in influenced by physiological and psychological causes. Obviously it is desirable that such a method be supplanted by a machine test which will eliminate the variabilities in the sound source and in the human ear. Up to the present time the nature of speech and the characteristics of the human ear are not known sufficiently well to establish either an ideal sound source or an electrical meter to replace the human voice and ear respectively. The best that can be done is to approximate this condition. Even though the meter readings may be the same, the simultaneous observations made with the ear in general will be different. A calibration of the machine must, therefore, depend upon a study of the degree of correlation between the average measure given by the machine and that given by the older method of test.

Thus, we see how special problems arise in the fields of both physics and engineering wherein it is impossible to control the variations. In what way, if any, are these problems related, or is it necessary to attack each one in a different manner? We shall see that all of these problems are in a way fundamentally the same and that the same method of solution can be applied to all of them. This is true because it is necessary to determine in every instance the law of distribution of the variable about some mean value.

Why Do We Need to Know the Law of Deviation of the Different Observations About Some Mean Value?

In all of the above problems as in every physical and engineering one, certain typical questions arise which can be answered only if we know the law of distribution $y=f(x)$ of the observations where y represents the frequency of occurrence of the deviations x from some mean value. At least three of these questions are the same for both fields of investigation.[4]

Let us consider the physical problem. From a group of n observations of the magnitude of a physical quantity, we obtain in general n distinct values which can be represented by $X_1, X_2, \ldots X_n$. From a study of these we must answer the following questions:

1. What is the most probable value?
2. What is the frequency of occurrence of values within any two limits?
3. Is the set of observations consistent with the assumption of a random system of causes?

The answers to these questions are necessary for the interpretation of Prof. Rutherford's data referred to above: They are required in order to interpret the data presented in Fig. 1 which are typical of physical and chemical problems arising in carbon study; these same answers are fundamentally required in the analysis of all physical data. These questions can be answered from a study of the frequency distribution. If this be true, it is obvious that the statistical methods of finding the best distribution are of interest to the physicist.

Let us next consider the engineering problem where we shall see that the same questions recur. Assuming that manufacturing methods are established to produce a definite number of instruments within a fixed period, one or more of the characteristics of these instruments must be controlled. We may represent any one of these characteristics by the symbol X. The total number of instruments that will be manufactured is usually very indefinite. It is, however, always finite. Even with extreme care some variations in the methods of manufacture may be expected which will produce

[4] In order to calibrate the machine referred to in a preceding paragraph and also to determine the relationships between the physico-chemical and microphonic properties of carbon, it was necessary to study the correlation between two or more variables, but in each case it was necessary to determine first the law of distribution for each variable in order to interpret the physical significance of the measures of correlation because this depends upon the laws of distribution. The reason for this is not discussed in the present paper, for attention is here confined to the method of establishing the best theoretical frequency distribution derived from a study of the observations.

variations from instrument to instrument in the quantity X. After the manufacturing methods have been established, the first problem is to obtain answers to the following questions:

1. What is the most probable value of X?
2. What is the percentage of instruments having values of X between any two limits?
3. Are the causes controlling the product random, or are they correlated?[5]

In this practical case we must decide to choose a certain number of instruments in order to obtain the answers to these questions; that is, to obtain the most probable frequency distribution. We must, however, go one step further. We must choose a certain number of instruments at stated periods in order to determine whether or not the product is changing. How big a sample shall we choose in the first place, and how large shall the periodic samples be? Obviously it is of great economic importance to keep the sample number in any case at a minimum required to establish within the required degree of precision the answers to the questions raised.

The close similarity between the physical and engineering problems must be obvious. Naturally, then, we need not confine ourselves in the present discussion to a consideration of only the problems arising in connection with the study of those microphonic properties of carbon which gave rise to the present investigation. Several examples are therefore chosen from fields other than carbon study. However only those points which have been found of practical advantage in connection with the analysis of more than 500,000 observations will be considered.

The type of inspection problem may be illustrated by the data given in Table II.

The symbol X refers to the efficiency of transmitters as determined in the process of inspection: N represents the number of instruments measured in order to obtain the average value $\bar{X}$. The first four rows of data represent the results obtained by four inspection groups G_1, G_2, G_3 and G_4. The results given are for the same period of time. The next three rows are those for different machines M_1, M_2 and M_3. The last row gives the results of single tests on 68,502 transmitters, a part of which was measured on each of the three machines. The third column in the table gives the standard deviations. It will be observed

[5] The significance of this question will become more evident in the course of the paper. We shall find that, if the causes are such as to be technically termed random, we can answer all practical questions with a far greater degree of precision than we can if the causes are not random.

TABLE II

INSPECTION DATA ON TRANSMITTERS

	$\bar{X}$	σ	$3\sigma/\sqrt{N}$	N	k	σ_k	β_2	σ_{β_2}	$\sigma_{\bar{X}}$	$3\sigma_k$	$3\sigma_{\beta_2}$	Pearson Type
G_1	548	739	0131	4510	214	0564	152	073	011	108	219	IV
G_2	740	896	0533	2540	949	0194	426	097	018	147	291	VI
G_3	766	762	0568	1620	109	0615	176	122	019	183	366	VI
G_4	934	677	0398	2610	1 413	0487	677	096	013	144	288	IV
M_1	−1 66	1 32	0386	10855	− 70	0243	128	047	013	072	141	
M_2	−1 69	1 07	0300	11577	− 84	0234	240	046	010	069	138	
M_3	−1 79	1 04	0510	3749	− 56	0403	628	080	017	120	240	
Machines 1, 2, 3	−1 641	1 14	0131	68502	− 80	0090	Out		004	027		I

that comparatively large differences exist between the averages obtained for different groups of transmitters by different groups of observers. Similarly, comparatively large variations exist in these averages even when taken by the machines (the large difference between the sensation and machine measures is due to a difference in the standard used, corrections for which are not made in this table).

Are these differences significant? Is product changing? That is, are the manufacturing methods being adequately controlled? Are these results consistent with a random variation in the causes controlling manufacture? These are the questions that were raised in connection with the interpretation of these data. The ordinary theory of errors gives us the following answer. It will be recalled that the standard deviation (or the root mean square deviation) of the average $\sigma_{\bar{X}}$ is equal to $\frac{\sigma}{\sqrt{N}}$. Also, from the table of the normal probability integral we find that the fractional parts of the area within certain ranges are as follows: For the ranges $\bar{X}\pm\sigma$, $\bar{X}\pm 2\sigma$, and $\bar{X}\pm 3\sigma$, we have the percentages 68.268, 95.450, and 99.730 respectively. Obviously, it is highly improbable that the difference between averages should be greater than three times the standard deviation of the average, providing we assume that all of the samples were drawn from the same universe: In other words, that all of the samples were manufactured under the same random conditions. The fourth column, then, indicates practical limits to the variations in the averages. It is obvious, therefore, that the differences between the averages are larger than could have been expected, if the same system of causes controlled the different groups of observations. In other words the differences are significant and must be explained.

Why do these variations exist? We shall show in the course of the discussion that the normal law is not sufficient to answer these questions. We shall show also that the variations noted are largely the result of the method of sampling used at that time. The significance of the other factors given in this table is discussed later.

Why Is the Application of the Normal Law Limited?

Why can we not assume that the deviations follow the normal law of error? This is

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \tag{1}$$

where σ is the root mean square error $\sqrt{\frac{\Sigma yx^2}{n}}$ and y is the frequency of occurrence of the deviation x from the arithmetic mean and n is the number of observations? If they do, the answers to all of the questions raised in the preceding paragraphs can be easily answered in a way which is familiar to all acquainted with the ordinary theory of errors and the method of least squares. This is an old and much debated question in the realm of statistics. Let us review briefly some of the a posteriori and a priori reasons why the normal law has gained such favor and yet why it is one of the most limited, instead of the most general, of the possible laws.

A Posteriori Reasons. The original method of explaining the normal law rests upon the assumption that the arithmetic mean value of the observations is always the most probable. Since experience shows that the observed arithmetic mean seldom satisfies the condition of being the most probable we may justly question the law based upon an apparently unjustified assumption.

Gauss first enunciated this law which is often called by his name. The fact that so great a mathematician proposed it led many to accept it. He assumes that the frequency of occurrence of a given error is a function of the error. The probability that a given set of n observations will occur is the product of the probabilities of the n independent events. He then assumes that the arithmetic mean is the most probable and finds the equation of the normal law. Thus he *assumes* the answer to the first question; that is, he assumes that the most probable value is always the arithmetic mean. In most physical and engineering measurements the deviations from the arithmetic mean are small, and the number of observations is not sufficiently large to determine whether or not they are consistent

with the assumption of the normal law. Under these conditions this law is perhaps as good an approximation as any.

The fundamental assumptions underlying the original explanation were later brought into question. What a priori reason is there for assuming that the arithmetic mean is the most probable value? Why not choose some other mean?[6] Thus if we assume that the median[7] value is the most probable, we obtain as a special case the law of error represented by the following equation:

$$y = Ae^{-h^2 x} \tag{2}$$

where y represents the frequency of occurrence of the deviation x from the median value and e is the Naperian base of logarithms. Both A and h are constants. If, however, we assume that the geometric mean is the most probable, we have as a special case the law of error represented by the following equation:

$$y = Ae^{-h^2 (\log X - \log a)^2} \tag{3}$$

where in this case y is the frequency of occurrence of an observation of magnitude X, "a" is the true value, and A and h are constants.[8]

Enough has been said to indicate the significance of the assumption that the arithmetic mean is the most probable value, but, why choose this instead of some other mean? No satisfactory answer is available. So far as the author has been able to discover, no distribution representing physical data has even been found which approaches the median law. Several examples have been found in the study of carbon which conform to the law of error derived upon the assumption that the geometric mean is the most probable. If the arithmetic mean were observed to be the most probable in a majority of cases, we might consider this an a posteriori reason for accepting the normal law. We find the contrary to be the case.

Furthermore, we find in general that the distribution of errors is non-symmetrical about the mean value. In fact, most of the distributions which are given in textbooks dealing with the theory of errors and the method of least squares to illustrate the universality

[6] An average or mean value may be defined as a quantity derived from a given set of observations by a process such that if the observations became all equal, the average will coincide with the observations, and if the observations are not all equal, the average is greater than the least and less than the greatest.

[7] If a series of n observations are arranged in ascending order of magnitude, the median value is that corresponding to the observation occurring midway between the two ends of the series.

[8] A very interesting discussion of the various laws that may be obtained by assuming different mean values is given in J. M. Keynes' "A Treatise on the Theory of Probability."

of the law are, themselves, inconsistent with the assumption of such a law. Prof. Pearson was one of the first to point out this fact. He considers among others an example originally given by Merriman [9] in which the observed distribution is that of 1,000 shots fired at a target. The theoretical normal is the solid line in Fig. 5 and the

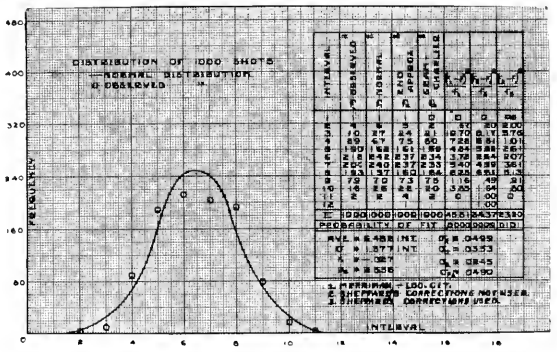


Fig. 5

observed frequencies are the small circles. When represented in this way there appears to be a wide divergence between theory and experience. Of course, some divergence may always be expected as a result of variations due to sampling; and, too, we must always question a judgment based entirely upon visual observation [10] of a graphical representation of this character. Prof. Pearson uses his method—which will be discussed later—for measuring the goodness of fit between the theoretical and observed distributions. He [11] finds that a fit as bad or worse than that observed could have been expected to occur on an average of only 15 to 16 times in ten million. We must conclude, therefore, that these data are not consistent with the assumption of a universal normal law.

A Priori Reasons. From the physicist's viewpoint the origin of the Gaussian law may be explained upon a more satisfactory basis.

[9] "Method of Least Squares," Eighth Edition—Page 14.

[10] This point will be emphasized later:- first, by showing that these data *appear* consistent with a normal law when plotted on probability paper, and second, by showing that some frequency distributions appear normal when plotted even though they are not. The other data in this table will be referred to later.

[11] Reference to the original article and a quotation therefrom given in the eleventh edition of the *Encyclopedia Britannica* on the article "Probability."

It is that which was originally suggested by La Place. If, however, we accept this explanation, we must accept the fact that the normal law is the exception and not the rule. Let us consider why this is true.[12]

This method of explanation rests upon the assumption that the normal law is the first approximation to the frequencies with which different values will be assumed by a variable quantity whose variations are controlled by a large number of independent causes acting in random fashion. Let us assume that:

a. The resultant variation is produced by n causes.
b. The probability p that a single cause will produce an effect Δx is the same for all of the causes.
c. The effect Δx is the same for all of the causes.
d. The causes operate independently one of the other.

Under these assumptions the frequency distribution of deviations of $0, 1, 2 \ldots n$ positive increments can be represented by the successive terms of the point binomial $N(q+p)^n$ where N represents the total number of observations.

Under these conditions if $p=q$ and $n=\infty$, the ordinates of the binominal expansion can be closely approximated by a *normal* curve having the same standard deviation. These restrictions are indeed narrow. In practice it is probable that p is never equal to q, and it is certain that n is never infinite. Therefore, the normal distribution should be the exception and not the rule.

There is a more fundamental reason, however, why we should seldom expect to find an observed distribution which is consistent with the normal law. In what has preceded we have assumed that each cause produced the same effect Δx, and that the total effect in any instance is proportional to the number of successes.

Let us assume that the resultant effect is, in general, a function of the number n of causes producing positive effects, that is, let $X=\phi(n)$. Thus we assume that the frequency distributions of the number of causes and of the occurrence of a magnitude X are respectively

$$y=f(n)$$

and

$$y_1=f_1(X)$$

for two values of n, say n and $n+dn$, there will be two values of X, say X and $X+dX$. The number of observations within this interval of n must be the same as that within the corresponding interval of X.

[12] Bowley "Elements of Statistics," Part II.

If the distribution in X is normal such that we have

$$y_1 = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-a)^2}{2\sigma^2}},$$

then

$$y = \frac{1}{\sigma\sqrt{2\pi}} \phi'(u) e^{-\frac{[\phi(u)-a]^2}{2\sigma^2}} \tag{4}$$

where a is the arithmetic mean value, therefore, the distribution of the causes need not be normal; conversely if the causes are distributed normally, the observations will not in general be normal.

This idea is of great importance in the interpretation of observed distributions of physical data.[13] To illustrate, let us assume that the natural causes which affect the growth of apples on a given tree produce a normal variation in the diameters of the apples. Obviously, the distribution of either the cross-sectional areas or the volumes will not be normal.[14]. If the distribution of the diameters is normal as supposed, the arithmetic means of these diameters is the most probable value. Obviously, however, neither the arithmetic mean area nor the arithmetic mean volume will be the most probable, because in general

$$\frac{1}{n}\Sigma f(X) \neq f\left(\frac{1}{n}\Sigma X\right). \tag{5}$$

As already indicated, the deviations dealt with in the present investigation were not small. The form of the observed distribution may be expected, therefore, to depend upon the functional relationship between the observed quantity and the number of causes. We shall

[13] Kapteyn, J. C.—Skew Frequency Curves—Groningen, 1903.

[14] In the theory of errors this fact is taken into account by assuming that the variations are always *small*. Thus, if the variable X can be represented as a function F of certain other variables U_1, U_2, ... U_m so that we have

$$X = F(U_1, U_2, \ldots U_m),$$

we ordinarily assume that we can write this expression in the following form

$$X = F(a_1 + u_1, a_2 + u_2, \ldots a_m + u_m).$$

A further assumption is made that the u's are small so that 2nd and higher powers and products of these can be neglected. Under these conditions the distribution of X is normal and has a standard deviation given by the following expression:

$$\sigma_X = \sqrt{\left(\sigma_{U_1}\frac{\partial F}{\partial U_1}\right)^2 + \left(\sigma_{U_2}\frac{\partial F}{\partial U_2}\right)^2 + \ldots \left(\sigma_{U_m}\frac{\partial F}{\partial U_m}\right)^2}.$$

But, thus, we are led to overlook the significance of the form of F, particularly in those practical cases such as are of interest in the present paper where the quantities $u_1, u_2, \ldots u_m$ are not small.

illustrate the significance of these ideas as an aid in the interpretation of data by reference to the results of our study of the law of error of the human ear in measuring the efficiency of transmitters.

Let us consider the problem of determining the minimum audible sound intensity. Let us assume that there are n physiological and psychological causes controlling this sensation measure, and that the probabilities of the causes producing 0, 1, 2 . . . n effects are dis-

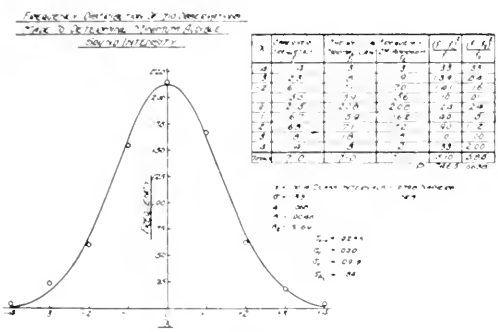


Fig. 6

tributed normally. Because of these differences in human ears different amounts of sound energy are required to produce minimum audible sensations. What is the distribution of energies?

The data are given in Fig. 6. These have been previously reported by Fletcher and Wegel of this laboratory.[15] The method of making these measurements was described in their original papers. It is sufficient to recall that the results are given in terms of pressures in dynes per square centimeter. Seven hundred and ten observations covering the frequency range of from 60 to 8,000 cycles are included. The data include results for both ears of 11 women and 20 men, and one ear only for two women and two men. Only ears that had been medically inspected as being physiologically normal were selected. These results, therefore, include variations in the observations of a single observer with those of different observers.

The natural logarithms of the intensities were added and the average of these was obtained. The distribution of the natural

[15] Fletcher, H. and Wegel, R. L., Proceedings of the National Academy of Science, Vol. VIII, pp. 5–6, January, 1922
Physical Review, Vol. XIX, pp. 550 seq. 1922

logarithms of the intensities is given in the second column of the table in Fig. 6. The smooth line is the normal curve based upon the observed value of standard deviation. The distribution of the logarithms of the intensities is normal.[16] The arithmetic mean of the logarithms is the most probable. Therefore, the distribution of intensities is decidedly skew, and the geometric mean intensity is the most probable. Here, then, is an excellent example in which it is highly probable that the distribution of the causes is random and normal, but in which the resultant effect is not a linear function of the number of causes.[17]

Can We Ever Expect to Find a Normal Distribution in Nature?

The answer is affirmative. If the resultant effect of the independent causes is proportional to their number, the distribution rapidly approaches normality as the number of causes is increased even though $p \neq q$.

To show this, let us assume that the variation in a physical quantity is produced by 100 causes, and that each cause produces the same effect Δx. Also, let us assume the probability p to be 0.1, that each cause produces a positive effect. The distribution of $0, 1, 2, \ldots n$ successes in 1000 trials is given by the terms of the expansion $1000\,(.9+.1)^{100}$. Obviously such a distribution is skew, p is certainly not equal to q, and n is far from being infinite. If the normal law

[16] In fact this is an exceptionally close approximation to the normal law. This will be more evident after we have considered the methods for measuring the goodness of fit as indicated by the other calculations given in this figure. For the present it is sufficient to know that approximately 75 times out of 100 we must expect to get a system of observations which differ as much or more from the theoretical distribution calculated from the normal law than the observed distribution differs therefrom in this case. The fact that the second approximation does not fit the observed distribution as well as the normal—*i.e.* the measure of probability of fit P is less—indicates that the observed value of the skewness k is not significant.

[17] These results are of particular interest to telephone engineers. The fact that the distribution of the logarithms of the intensities is normal is consistent with the assumption of Fechner's law which states that the sensation is proportional to the logarithm of the stimulus. The range of variation (that is, $X = 3\sigma$) in different observers' estimates of the sound intensity required to produce the minimum audible sensation is approximately 20 miles. The range of error of estimate depends upon the intensity of sound and decreases as the sound energy level increases. Thus for the average level which prevails for transmission over the present form of telephone system in a three mile loop common battery circuit it is less than 9 miles. Even at this intensity, however, it is obvious that although scarcely any observers will differ in their estimates by more than 9 miles, 50% of them will differ by at least 2 miles. These results also furnish experimental basis for the statement made in the beginning of this paper: that is, the variations introduced in the method of measurement of transmitter efficiencies are large in comparison with the average efficiency.

TABLE III.—COMPARISON OF THE THREE AND FOUR-FOLD APPROXIMATIONS

Number of Successes	* $1000(.1+.9)^{100}$ f	Normal Law f_1	2nd Approximation f_2	*Law of Small Numbers f_3	*Gram-Charlier f_4	*Poisson Charlier f_5	†Pearson f_6	$(f-f_1)^2$ f_1	$(f-f_2)^2$ f_2	$(f-f_3)^2$ f_3	$(f-f_4)^2$ f_4	$(f-f_5)^2$ f_5	$(f-f_6)^2$ f_6
−1	0	2	.0	0	0	0	.0	.200	0	.0	0	.0	0
0	1	6	0	.0	.0	0	.9	.417	∞	∞	∞	∞	.711
1	3	1 5	.4	5	4	2	2 2	960	025	080	025	050	1 641
2	1 6	3 9	2 1	2 3	2 0	1.6	7 1	1.356	.119	213	080	0	4 261
3	5 9	8 9	6 7	7 6	6 5	5 9	15.9	1 011	.096	380	055	0	6 289
4	15 9	18 3	16 7	18 9	16 2	15 9	28.9	315	038	476	006	0	5 848
5	33 9	33 4	34 0	37 8	33 3	34 0	44 8	007	.0	492	011	0	2 652
6	59 6	54 8	58 8	63 0	58 1	59 9	61 3	420	011	183	039	002	047
7	88 9	80 9	88 0	90 1	87 5	89 3	76 1	791	009	016	022	002	2 153
8	114 8	106 0	113 9	112 5	114 5	114 9	87 0	731	007	047	001	0	8 883
9	130 4	125 2	130 5	125 1	131 8	130 1	92 8	216	0	225	015	001	15 234
10	131 9	132 6	132 6	125 1	133 9	131 4	93 3	004	004	370	030	002	15 970
11	119 9	125 2	119 9	113 7	121 1	119 4	89 1	224	0	338	012	002	10 647
12	98 8	106 0	98 1	91 8	98 4	98 6	81 2	489	005	169	002	0	3 815
13	71 3	80 9	73 8	72 9	73 2	74 4	71 0	538	003	027	017	0	153
14	51 3	54 8	50 8	52 1	50 2	51 4	59 9	224	005	012	024	0	1 235
15	32 7	33 4	32 8	31 7	32 2	33 0	48 7	015	0	115	008	003	5 257
16	19 3	18 3	19 9	21 7	19 4	19 7	34 5	055	018	265	001	008	6 697
17	10 6	8 9	11 1	12 8	10 9	10 8	29 5	325	023	378	008	004	12 109
18	5 4	3 9	5 7	7 1	5 8	5 5	22 0	577	016	407	028	002	12 525
19	2 6	1 5	2 6	3 7	2 7	2 5	16 0	807	0	327	004	004	11 223
20	1 2	6	1 3	1 9	1 2	1 1	11 4	600	008	258	0	009	9 126
21	5	1	4	9	.5	.5	7.9	1 600	025	178	.0	0	6 932
22	2	1	.2	4	.2	.1	5.4	.100	.0	100	0	1	5 007
23	0	0	.0	2	0	.0	3 6			200			3 600
24	0	0	.0	.0	0	.0	2 3						2 300
25	0	0	0	.0	.0	.0	1 5						1 500
26	0	0	0	.0	.0	0	.9						900
27	0	0	0	.0	.0	0	.6						600
28	0	0	0	.0	.0	0	4						400
29	0	0	0	0	0	0	.2						200
30	0	0	0	.0	.0	0	.1						100
Σ	1000 1	1000 0	1000 3	999 8	1000 0	1000 2	996 5	11.982	112	5 166	388	189	158 015
Ave.	9 998	10 000	9 996	9 999	9 998	10 000	10 754						

$$p = 1 \qquad \sigma = \sqrt{pqn} = 3 \qquad \beta_2 = 3 + \frac{1 - 6\ pq}{pqn} = 3\ 511 \qquad \sigma_{\beta_2} = \sqrt{\frac{24}{1000}} = 155$$

$$\begin{matrix} q & 9 \\ n & 100 \\ pn & 10 \end{matrix} \qquad K = \frac{q - p}{\sqrt{pqn}} = .267 \qquad \sigma_K = \sqrt{\frac{6}{1000}} = .077$$

* Fisher, A. "The Mathematical Theory of Probabilities."

† Pearson, Karl. Phil. Mag., 1907, pp. 365–378.

were fitted to such a distribution, would it be possible to detect easily any great difference between theory and observation?

Let us compare the two distributions. The data are given in Table III. First, the average value must be the most probable in order to be consistent with the normal law. It is, because the observed most probable value corresponds to 10 successes, and the average of

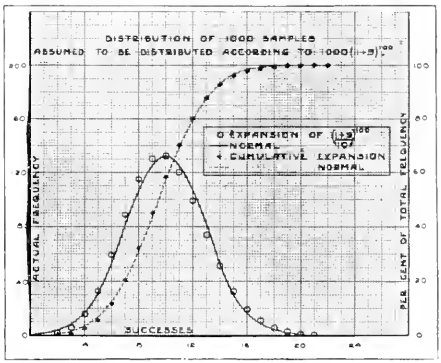


Fig. 7

the hypothetically observed distribution is 9.998. This under ordinary circumstances would be considered a close check between theory and practice.

The normal distribution is given in the third column of the table. Even though there is a difference between the frequencies given in the second and third columns, would the average observer be apt to conclude that the hypothetically observed distribution is other than normal? He would probably base his answer upon a graphical comparison such as given in Fig. 7. The solid line represents the normal curve; whereas the frequencies given in the second column of Table III are represented by circles. It is obvious that the normal law appears to be a very close approximation to the terms of the binomial expansion.

Thus we see that for even a small number of causes the difference between p and q may be quite large, and yet the difference between the distributions given by the binomial expansion and that given by the normal law is apparently small and not easily to be detected by ordinary methods. As n increases the closeness of fit does likewise.

If p is equal to q, the number of causes must be very small indeed before we are able to detect the difference between the terms of the binomial expansion and those given by the normal law. To show that this is true I have chosen a case corresponding to a physical condition where there are only 16 causes and where p is equal to q. The data are given in Table IV.

TABLE IV

Successes	$(\frac{1}{2} + \frac{1}{2})^{16}$ f	Normal Law with same σ f_1
0	0000153	0000669
1	0002441	0004363
2	0018311	0022159
3	0085449	0087641
4	0277710	0269955
5	0666504	0647586
6	1220825	1209853
7	1745605	1760326
8	1963806	1994711
9	1745605	1760326
10	1220825	1209853
11	0666504	0647588
12	0277710	0269955
13	0085449	0087641
14	0018311	0022159
15	0002441	0004363
16	0000153	0000669

Obviously, therefore, the limitations imposed by the assumptions as to the number of causes and the equality of p and q are not as important as they might at first appear. It is probable that this is one of the reasons why we find approximately normal distributions. If, however, p is sufficiently small, the difference between the observed distribution and that consistent with the normal law can easily be detected. We shall show in a later section that this is true for Rutherford's data.[18]

Is There a Universal Law of Error?

Obviously from what has already been said, the normal law is not a universal law of nature. It is probable that no such law exists. We do, however, have certain laws which are more general than the normal. We shall consider briefly some of these types in an effort to indicate the advantages that can be gained by an application of them to physical data.

[18] Loc. cit.

Binomial Expansion $(p+q)^n$. We have already seen that the distribution is approximately normal when $p=q$ and $n=\infty$. Following Edgeworth[19], Bowley [20] shows that if $p\neq q$ but $n=\infty$ the frequency y of the occurrence of a deviation of magnitude x is given by the following expression where k represents the skewness [21] of the distribution:

$$y=\frac{1}{\sigma\sqrt{2\pi}}\left(\exp.-\frac{x^2}{2\sigma^2}\right)\left[1-\frac{k}{2}\left(\frac{x}{\sigma}-\frac{x^3}{3\sigma^3}\right)\right]. \tag{6}$$

This will be referred to as the *second approximation*.

If p is very small, but $pn=\lambda$ is finite, we have the so-called *law of small numbers* [22] which was first derived by Poisson. The successive terms of the series $e^{-\lambda}\left(1+\lambda+\frac{\lambda^2}{2}+\frac{\lambda^3}{3}+\ldots\right)$ represent the chances of 0, 1, 2 . . . n successes. Theoretically, if we are dealing with a distribution of attributes,[23] it is always possible to calculate the values of

[19] Cambridge Philosophical Transactions, Vol. XX, 1904, pp. 36-65 and 113-141.

[20] Loc. cit.

[21] In statistical work the practice is followed of using the moments of the distribution for determining the parameters of the frequency curve. The i th moment μ_i of a frequency distribution about the arithmetic mean is by definition

$$\mu_i=\frac{\Sigma yx^i}{\Sigma y}$$

In calculating such moments it is necessary to consider the observations as grouped about the mid-point of the class interval and unless this interval is very small certain errors are introduced which can be partially eliminated by applying Sheppard's corrections as given by him in Biometrika, Vol. III, pages 308 seq. If Δx be taken as unity, we have

$$\mu_1=0$$
$$\mu_2=pqn=\sigma^2$$
$$\mu_3=pqn(q-p)$$
$$\mu_4=3(pqn)^2+pnq(1-6pq)$$

$$\beta_1{}^2=k=\frac{q-p}{\sqrt{pqn}}$$
$$\beta_2=3+\frac{1-6pq}{pqn}$$

and if p is approximately equal to q and n is large we have $\sigma_k=\sqrt{\frac{6}{N}}$ and $\sigma_{\beta_2}=\sqrt{\frac{24}{N}}$.

[22] It is of interest to note that several investigators have derived this law independently. Thus H. Bateman derives this expression in an appendix to the article of Prof. Rutherford and H. Geiger previously referred to. This is, in a way, an illustration of the apparent need of a broader dissemination of information relating to the application of statistical methods of analysis to engineering and physical data. It is also of interest to note that this law has been used to advantage in the discussion of telephone trunking problems.

[23] If the classification is based upon the presence or absence of a single characteristic, this characteristic is often referred to as an attribute.

p, q and n from the moments of the distribution.[24] Even when p, q and n are known, the arithmetic involved in calculating the terms of the binomial is often prohibitive, and, therefore, it is necessary to obtain certain approximations corresponding to the three laws of error; that is, normal, second approximation, and the law of small numbers. Tables for the normal law and for the law of small numbers are readily available in many places, while those for the second approximation are given by Bowley.[25]

Even under conditions where the binomial expansion does not hold, Edgeworth has shown that it is possible to obtain the following general approximation:

$$y = \frac{1}{\sigma\sqrt{2\pi}}\left(\exp.-\frac{x^2}{2\sigma^2}\right)\left[1-\frac{k}{2}\left(\frac{x}{\sigma}-\frac{x^3}{3\sigma^3}\right)+\frac{k^2}{8}\left(-\frac{5}{3}+\frac{5x^2}{\sigma^2}-\frac{5x^4}{3\sigma^2}+\frac{x^6}{9\sigma^6}\right)+\left(\frac{\mu_1-3\mu^2}{8\sigma^4}\right)\left(1-\frac{2x^2}{\sigma^2}+\frac{x^4}{3\sigma^4}\right)\right]. \quad (7)$$

This holds providing the observations are influenced by a large number of causes, each of which varies according to some law of error but not necessarily to the normal law.

Gram-Charlier Series. Gram, according to Fisher,[26] was the first to show that the normal law is a special case of a more generalized system of skew frequency curves. He showed that the arbitrary frequency function $F(X)$ can be represented by a series of terms in which the normal law is the generating function $\phi(X)$. Thus

$$F(X) = c_0\phi(X)+c_1\phi'(X)+c_2\phi''(X)+\ldots \quad (8)$$

where c_0, c_1, c_2, etc., are constants which may be determined from the moments of the observed data. This series is similar to that already mentioned in the above equation (7) which Edgeworth has obtained in several different ways. This law is of interest from the viewpoint of either a physicist or an engineer in so far as it gives him a picture of the casual conditions consistent with an accepted theoretical curve. Thus, if either the causes of variation are within a certain degree not entirely independent, or the errors are not linearly aggregated, the observed frequency distributions may be expected to conform to an equation such as 8. This equation has been found to fit a much larger group of observed distributions than the normal law

[24] See footnote 26.

[25] See for example Pearson, K.—Tables for Biometricians and Statisticians—Cambridge University Press.

[26] Fisher, Arne—Theory of Probabilities—page 182.

and the publication of the necessary tables by Fisher [27] and Glover [28] makes the study of such a curve more feasible. The author finds, for example, that this series furnishes a much closer fit to the distribution of shots, Fig. 5, referred to above than any other that he has tried.

Theoretically we should be able to improve the approximation by taking a large number of terms of the series. Such a procedure, however, involves the use of moments higher than the first four, and the errors in these moments are so large as to make their use impractical.

In spite of the uncertainty attached to the interpretation of the physical significance of fitting any of these curves to data, one very practical observation has been made: that is, if an observed series of frequencies could not be fitted by a theoretical curve in any of the ways already mentioned, careful consideration of the possible reasons for the observed poor fit have in practically every instance suggested the cause or causes thereof. We shall refer to only one practical example.

The data have already been given above in Table II. It has been noted that in this instance the variations in the averages of groups of several thousand observations showed that the differences were significant. If the observed distributions had been normal, it would have been necessary to assume either that the methods of making the measurements were different for the different groups of observers, and for the different machines, or that the manufacturing methods were experiencing a trend. Although the observed frequency curves for the different groups were found to be smooth, the observed frequencies could not be readily fitted by any curve previously described. This naturally led to a search for the existence of any one of a number of causes affecting the observations which might produce such a divergence between theory and practice. One by one these causes were found and eliminated and as they were the degree of fit between the results of theory and practice increased. For example, it was found that some of the groups of observations were for transmitters assembled from only two or three lots of carbon. Transmitters assembled from one lot of carbon had a different average efficiency from those assembled from another lot. Naturally the

[27] Fisher, Arne Loc. cit. As noted by Mr. Fisher, page 214, the values of $\phi(x)$ and its first 6 derivatives to 7 decimal places for values of x up to 4 and progressing by intervals of 0.01 were given by Jorgensen in his "Frekvensflader og Korrelation."

[28] Glover, J. W. Tables of Compound Interest, Functions, etc. 1923 Edition published by George Wahr, Ann Arbor, Michigan.

resultant distribution was a compound of a few separate but similar distributions about different averages. When the distributions of the efficiencies of the different lots of carbon were determined separately they were found to be consistent with the second approximation.

Thus, although it may be impossible to conclude that the a priori assumptions underlying a given law of distribution are fulfilled because the observations are found to be consistent therewith, nevertheless, the fact that the observed and the theoretical distributions do not agree suggests the necessity of seeking for certain typical causes which may be expected to introduce such discrepancies. This point is of special importance in connection with the study of ways of sampling product in order to determine whether or not the manufacturing process is subject to trends. Thus, if a product is sampled at two periods, and the distributions of both groups of observations are found to be random about different averages, it is highly probable that the difference indicates a trend in the manufacturing methods, providing the difference between the averages is greater than 3 times the standard deviation of the average. When, however, the two distributions are found to be inconsistent with a random system of causes, it is quite probable that the condition of sampling has not been carefully controlled.

Hypergeometric Series. Pearson has shown several ways in which a frequency distribution may be represented by a hypergeometric series. Thus the chances of getting r, $r=1, \ldots 0$ bad transmitters from a lot containing pn bad and qn good and where r instruments are drawn at a time may be represented by the terms of such a series. More important, however, is Pearson's solution [29] of what he calls the fundamental problem of statistics. He shows, following the line of reasoning similar to that originally suggested by Bayes, that if in a sample of $k_1=(m+n)$ trials, an event has been observed to occur m times and to fail n times, in a second group of k_2 trials the chances of the event occurring r times and failing s times are given by the successive terms of a hypergeometric series. We cannot consider here the questions underlying the justification of this method of solution, for, as is well-known, the application of Bayes' theorem is questioned by many statisticians. We can profit, however, by the broad experience of Prof. Pearson, for he has apparently accumulated an abundance of data which are consistent with the theory.

The answer to this problem is of special importance in connection with the inspection of product which in many instances runs into millions yearly. We must keep the cost of inspection at a minimum,

[29] Pearson, K. - Biometrika, October, 1920 -pp. 1-16.

which means that the sample numbers must be small, and yet we see from the solution derived from Pearson the significance of the sizes of both the original and the second sample. Thus, he [30] shows that the standard deviation σ is given by the equation

$$\sigma^2 = k_2 pq\left(1 + \frac{k_2}{k_1}\right). \tag{9}$$

Multimodal Distributions. These occur frequently in engineering work and particularly in connection with the inspection of large quantities of apparatus. One such instance has already been referred to in the discussion of the data given in Table II, and another is illustrated by the data given in Fig. 1. Prof. Pearson [31] has developed a method for determining analytically whether or not the observed distribution is such as may be expected to have arisen from the combination of two normal components, the mean values of which are different. The method involves the solution of a ninth degree equation. As a result, the arithmetic work is in many cases prohibitive. This method cannot be applied to the data given in Fig. 1 primarily because the number of observations is not sufficiently great.

Pearson's Closed Type Curves.[32] One of the best known statistical methods for graduating data is that developed by Prof. Pearson. His system of closed type curves arises from the solution of the differential equation derived upon the assumption that the distribution is uni-modal and touches the axis when $y = 0$. In the hands of Pearson and his school great success has been attained in graduating data collected from widely different fields, although primarily from these of biology, psychology, and economics. The choice of curve to represent a given distribution rests primarily upon a consideration of a criterion involving two constants, $\beta_1 = \sqrt{k}$ and β_2, both of which have been defined previously in footnote 21.

In the early study of the distributions of efficiencies of product transmitters an attempt was made to apply this system of curves. For example, the Pearson types are indicated in Table II. In no instance, however, was it possible to obtain a very satisfactory fit between the observed and the theoretical distributions. Furthermore, the arithmetical work required to calculate a theoretical distribution in this way is excessive. We must also consider what physical significance can be attached to the different types of curves. The answer is not definite. Under certain conditions the generalized

[30] Pearson, K.—*Philosophical Magazine*—1907, pp. 365-378.

[31] Pearson, K.—*Philosophical Magazine*—Vol. 1, 1901, pp. 115-119.

[32] Elderton—Frequency Curves and Correlation.

equation of Pearson breaks down to the normal law and the second approximation. These, of course, can be explained as previously. The fundamental equation, however, serves to cover the condition where the causes are correlated. Thus, because of the lack of a clear conception of the physical significance of the observed variations in the type of curves indicated in Table II, it was not possible easily to set up experiments to find the causes of these variations. For this reason preference has been given to the use of frequency distributions derived upon a less empirical basis following the original lines laid down by La Place, Edgeworth, Kapteyn, and others previously referred to. Another very practical reason for choosing the latter type of curve is that it involves for the most part the use of only the first three moments of the distribution instead of the first four required for differentiating between the Pearson types. In those cases where the interest is less of physical interpretation than of graduating an observed set of data, preference may go to the more generalized system of Pearson.

How Can We Choose the Best Theoretical Frequency Distribution?

We have already briefly reviewed some of the different methods for obtaining a theoretical frequency distribution from a consideration of the moments of the observed frequencies. We have seen in Table III that by using different methods we obtain different degrees of approximation to the hypothetically observed distribution which in this case corresponds to the terms of the binomial expansion $1000(.1+.9)^{100}$. Similarly from Fig. 5 it is seen that the Gram-Charlier series is a much closer approximation to the observed distribution than that derived upon the assumption of the normal law. In any given case we are naturally confronted with the question: What is the best theoretical distribution? We shall consider four methods for obtaining an answer.

The oldest, simplest, and in many instances the most practical, is that of comparing graphically or in tabular form the theoretical distribution with the one observed. This method is, however, inaccurate and qualitative. It does not furnish us with a quantitative method of measuring the closeness of fit between theory and practice, and in certain instances it is absolutely misleading. It is of interest to see how all of these things can be truly said of one and the same method. The first two characteristics, that is, oldest and simplest, are perhaps readily granted. It remains to be pointed out more

definitely wherein the method is sadly deficient as a quantitative measure, and therefore often misleading; whereas in certain instances it may be, nevertheless, the only practical method that can be used.

Graphical Method. The graphical method itself may be subdivided into two parts. Let us consider first the plot of the observed and theoretical frequencies. As an example of the unsatisfactory nature of this form of comparison, it is of interest to consider certain data

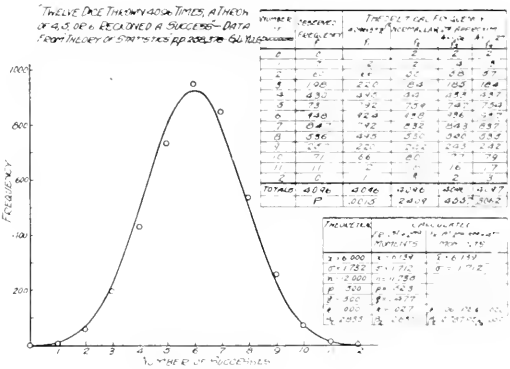


Fig. 8

given by Yule [33] in which 12 dice are thrown 4,096 times, a throw of 4, 5, or 6 points being reckoned a success. If the dice are symmetrical $p=q=\frac{1}{2}$ and the theoretical distribution if given by $4{,}096\,(\frac{1}{2}+\frac{1}{2})^{12}$, the terms of which as given by Yule are presented in the third column of Fig. 8. It is suggested that the reader, before going further, consider the graphical and tabular representation of these data. The smooth curve is the theoretical distribution $4{,}096(\frac{1}{2}+\frac{1}{2})^{12}$. It has been the author's experience to find that in practically every instance in which this curve has been shown to an individual for the first time that the impression is that which Yule evidently desires to produce by the illustration: that is there is a very good fit between theory and practice. This distribution is, however, not symmetrical: it is skew. The dice used in this experiment were not symmetrical: that is, $p \neq q$. How do we know that these statements are true?

Let us consider the normal and second approximation as given

[33] Yule—"Introduction to the Theory of Statistics."

in the fourth, fifth, and sixth columns.[34] Obviously the degree of fit is closest for the second approximation, although that between the normal distribution and the observed frequencies is closer than that between the terms of the binomial expansion and the observed frequencies. To be sure, the normal law is only an approximation to the point binomial when $p=q$ and $n=\infty$. The normal distribution, however, is calculated about the observed average 6.139, instead of about the theoretical average 6. If the dice are non-symmetrical, the average will not be 6, and, therefore, the center of the distribution will be shifted after the fashion observed. The improvement in fit corresponding to the normal distribution is therefore primarily attributable to that introduced by shifting the center of the distribution indicating that $p \neq q$. However, if $p \neq q$, the second approximation should improve the fit and for either value of k this is found to be the case. Thus even though we cannot measure quantitatively the improvement of fit, the qualitative evidence presented in this figure is sufficient to warrant the conclusion that the dice were non-symmetrical, and therefore, that the smooth curve is an unsatisfactory graduation of the data. In fact, by using a quantitative method for measuring the goodness of fit to be discussed in a succeeding paragraph, it follows that only 15 times out of 10,000 can we expect a divergence from theory as large or larger than that exhibited by the frequencies corresponding to the point binomial.

We have also previously called attention to the fact that in Fig. 7 the eye does not serve to differentiate satisfactorily between the distribution calculated upon the assumption of the normal law and that given by the binominal expansion when the conditions underlying the normal law are far from being satisfied.

Regardless of these criticisms, such graphical methods cannot be entirely dispensed with. Thus the graphical representation of the data given in Fig. 1 shows very clearly that the distribution is probably bimodal, although with no more observations than are available it is practically impossible to show that this is true in any other way.

Instead of plotting the frequency y of occurrence of a variable of magnitude x as ordinate, and x as abscissa, the practice is often followed of plotting as ordinate the percentage of the total number N of observations having magnitudes of x or less.[35]

Any curve $\phi(y, x)=0$ may be replaced by a straight line.[36] In

[34] Two values of k were calculated as indicated in the lower right hand corner of the figure.

[35] Heindlhoffer, K. and Sjovall, H. Endurance Test Data and their Interpretation. Advance paper presented at the Meeting of the American Society of Mechanical engineers, Montreal, Canada, May 28 to 31, 1923.

[36] Runge, C. *Graphical Methods*, p. 53.

this way we can transform the integral curve into a straight line by choosing an x-scale proportional to the integral from 0 to x of the probability curve.[37] When plotted in this way, a normal distribution appears as a straight line on such paper. At first it may appear very simple to determine whether or not the data conform to a straight line, but in practice this is not always so easy. Thus, we have seen that the distribution of shots presented in Fig. 5 is not normal, but

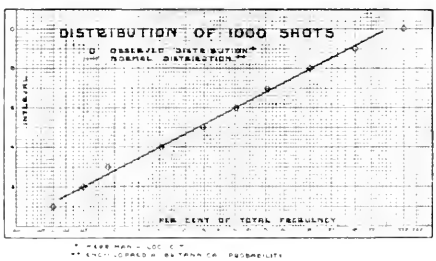


Fig. 9

when these results are plotted on probability paper we have the curve given in Fig. 9. The reader should be cautioned that in such a case there is a temptation to consider that the observed points are approximately well fitted by the straight line, although this is not the case.

Probability paper could be ruled for different theoretical distributions, but in its present form it serves only to determine whether or not the distribution is approximately normal. Its use leaves much to be desired in the way of a quantitative measure of the degree of fit between the theoretical and observed distributions.

Calculation of σ, $\beta_1 = \lambda k$, *and* β_2. Let us consider what information can be obtained as to the best theoretical distribution from only a consideration of the first four moments of the observed frequencies. Let us consider the values of k and β_2 presented in Table V. These have been calculated for the point binomial $(p+q)^n$ where p, q and n have been given different values. For the normal law corresponding to $p=q$ and $n=\infty$, we have $k=0$ and $\beta_2=3$. Thus, if in a practical

[37] Whipple, G. C.—The Elements of Chance in Sanitation—*Franklin Institute Journal*, Vol. 182, July, December, 1916 pp. 37-59 and 205-227.

TABLE V

p	$n=4$		$n=9$		$n=16$		$n=25$		$n=100$		$n=10{,}000$	
	k	β_2	k	β_2	k	β_2	k	β_2	k	β_2	k	β_2
.5	0	2.50	0	2.78	0	2.87	0	2.92	0	2.98	0	3.00
.6	−.20	2.51	−.14	2.80	−.10	2.89	−.08	2.93	−.04	2.98	−.004	3.00
.7	−.44	2.69	−.29	2.86	−.22	2.92	−.17	2.95	−.09	2.99	−.009	3.00
.8	−.75	3.06	−.50	3.03	−.38	3.02	−.30	3.01	−.15	3.00	−.015	3.00
.9	−1.33	4.28	−.69	3.57	−.67	3.32	−.53	3.20	−.27	3.05	−.027	3.00
.99	−4.92	26.75	−3.28	13.56	−2.46	8.94	−1.97	6.80	−.98	3.95	−.098	3.01
.999	−15.79	251.75	−10.52	113.45	−7.90	65.19	−6.31	42.76	−3.16	12.95	−.316	3.10
.9999	−50.00	2501.75	−33.33	1113.44	−25.00	627.69	−20.00	402.76	−10.00	102.94	−1.000	4.00

case we find an observed distribution for which $k=0$ and $\beta_2=3$, it is highly probable that the distribution is approximately normal. It is true, however, that in sampling from a universe in which $p=q$ and $n=\infty$, the observed values of k and β_2 will seldom be exactly equal to 0 and 3 respectively. Then we must ask what range of values may be expected in these two factors for distributions which are practically normal. For such cases the variations in k and β_2 are practically normal [38] and have standard deviations $\sigma_k=\sqrt{\frac{6}{N}}$ and $\sigma_{\beta_2}=\sqrt{\frac{24}{N}}$ where N is the number of observations. Thus, theoretically any series of observations for which the calculated values of k and β_2 fall within the ranges $0\pm3\sigma_k$ and $3\pm3\sigma_{\beta_2}$ may have arisen from a normal universe. Since, however, the errors σ_k and σ_{β_2} of sampling are so large, this method does not furnish a very practical test for distribution consisting of only a few observations. This is particularly true since, even for very skew distributions, the values of k and β_2 do not differ much from 0 and 3 respectively (see Table V). If, however, the number of observations is large, the values of k and β_2 in themselves often indicate very definitely that the observed frequencies are not consistent with the normal law. For example the calculated values of k and β_2 given for the inspection data in Table II show conclusively that in practically every instance the observed data could not have arisen from a normal universe. So long as we do not use Pearson's system of curves, all that these two factors indicate is that the observed data do or do not conform to the normal law and in this respect their use is limited as is that of the probability paper mentioned above.

In order to show that the factor β_2 is not in itself a very sensitive measure of the variability from the normal law, I have considered the following special case. Let us assume that the observed distributions can be grouped into two parts depending upon whether or not the observations cluster about the average X_1 or X_2 measured from a point which is the arithmetic mean of the entire distribution taken about a common origin. This corresponds to the practical case such as that indicated by Fig. 1 which as already pointed out often occurs in practice.

[38] For a critical study of the conditions under which the probable errors of these constants have a real significance, reference should be made to a discussion of this problem by Isserlis in the Proceedings of the Royal Society, series A, Vol. 92, pp. 23 seq. 1915. Obviously even for the normal distribution all of the moments will be skew. This follows from a consideration of equation 4.

The value of β_2 for the entire distribution is then given by the following expression:

$$\beta_2 = \frac{(N_1^4 \sum y_1 + N_2^4 \sum y_2) + 6(N_1^2\sigma_1^2 \sum y_1 + N_2^2\sigma_2^2 \sum y_2) + (N_1{}_1\mu_4 \sum y_1 + N_2{}_2\mu_4 \sum y_2)}{(\sum y_1 + \sum y_2)\mu_2^2},$$

where $_1\mu_i$ and $_2\mu_i$ refer to the adjusted ith moments of the observations about their respective mean values. Let us assume that $\overline{X} = X_1 = X_2$; $k_1 = k_2 = 0$; $_1\beta_2 = {}_2\beta_2 = 3$; $_1\mu_i = {}_2\mu_i$; $\sum y_1 = \sum y_2$; and $\sigma_1 = \sigma_2$ where $\sum y_1$ and $\sum y_2$ represent the total numbers of observations in the first and second groups respectively. It may be shown by substitution in this equation that, if $X_1 = \sigma_1$, $\beta_2 = 2.5$, whereas, if $X_1 = 10\sigma_1$, $\beta_2 = 1$, approximately. Thus, if the numbers of observations in each of the two sub-groups are the same and the component curves are normal, the value of β_2 for the entire distribution about the mean of the two will, in general, decrease as X_1 becomes large in comparison with σ_1. Differences in β_2 of this magnitude are difficult to establish. Furthermore the skewness is zero, and therefore does not indicate the bi-modal character of the distribution.

Let us consider the case where $a X_1 = X_2$; $k_1 = k_2 = 0$; $_1\beta_2 = {}_2\beta_2 = 3$; $\sum y_1 = a \sum y_2$; $_1\mu_i = {}_2\mu_i$. If, $a = 10$ and $X_1 = \sigma_1$ then $\beta_2 = 8+$ whereas if $X_1 = 10\sigma_1$, then $\beta_2 = 100$, approximately.[3] Thus, for comparatively wide differences in the averages, it requires a large number of observations in order to increase the precision of β_2 to such an extent as to prove the significance of deviations in this factor of the magnitudes noted above.

The skewness in this case is not zero and its significance could be established with a comparatively small number of measurements. In any of the above cases a carefully constructed plot would serve to indicate the bimodal characteristic of the curve better than the study of the factor β_2.

Pearson's Criterion of Goodness of Fit. A much more powerful

[3] Here again it should be noted that the values of β_2 are independent of the actual frequencies of each of the two groups and depend only upon the ratio of these frequencies and upon the ratio of X_1 to σ_1.

criterion has been developed by Prof. Pearson [40] in a series of articles in the *Philosophical Magazine*. It is true that this test for goodness of fit cannot be used indiscriminately. In fact the application of this criterion is subject to numerous limitations clearly set forth in the original papers by Pearson and in more recent articles on the mathematics of statistics. In the use of the method it is necessary that these be kept in mind by the individual making the original analysis of the data. Irrespective of these facts, however, the method itself is one of the most useful tools available for measuring in a quantitative way the "goodness of fit" between two distributions. The significance of the values of P given in Figs. 5, 6, and 8 now become evident.

Engineering Judgment. The fourth very practical and one of the most useful methods of comparing the theoretical with the observed distribution is that of applying common sense or engineering judgment. To quote from a recent article of Prof. Wilson [41] we have: "And as the use of the statistical method spreads we must and shall appreciate the fact that it, like other methods, is not a substitute for, but a humble aid to the formation of a scientific judgment." Even with the use of all the statistical methods known to the art, it remains impossible to determine the true nature of the complex of causes which control a set of observations. We can present plausible explanations, but we can never be sure that they are right. Sometimes we can present two plausible explanations and then we must fall back on engineering judgment or common sense to decide between them. A striking illustration of this fact is presented in the following paragraph.

Prof. Pearson [42] has recently presented measurements of the cephalic index of a certain group of skulls. The object of the investigation was to determine if variation had gone on to such an extent as to indicate the survival of the fitter inside a homogeneous population, or the survival of two races both of which were in existence many ages in the past. Pearson shows that, by a solution of a nonic equation,

[40] If we divide the entire range of variation into s equal intervals for which the observed frequencies are $f_1, f_2, \ldots, f_s$ and the corresponding theoretical frequencies are $f_1', f_2', \ldots, f_s'$, Pearson calculates the function

$$\chi^2 = \sum \frac{(f'-f)^2}{f'}$$

from which he is able to determine the probability that a series of deviations as large as, or larger than, that found to exist could have arisen as a result of random sampling. Tables have been prepared which give the probability of fit in terms of the number of intervals into which the entire range has been divided and of the value of χ.

[41] Wilson, E. B.- The Statistical Significance of Experimental Data –science – New Series, Vol. 58, 1493, October 10, 1923, pp. 93-100.

[42] *Philosophical Magazine*, Vol. 1, 1901 pp. 110-124.

he is able to find two component distributions which when added together approximate very closely to the observed frequencies. The observed data are given in the second column of Table VI and the frequencies of Prof Pearson's compound curve are given in the third column of the table. The probability of fit between these two distributions is seen to be approximately .96, which is indeed very

TABLE VI

ROMANO-AVAR SKULLS *

Cephalic Index	Observed Distribution f	Compound Distribution f_1	2nd Approximation f_2	$\frac{(f_1-f)^2}{f_1}$	$\frac{(f_2-f)^2}{f_2}$
67	1	1	1	0	0
68	1	2	2	.50	.50
69	3	4	4	.25	.25
70	8	7	8	.14	0
71	13	11	14	.36	.07
72	13	18	22	1.39	3.68
73	33	28	30	.89	.30
74	36	39	39	.23	.23
75	49	50	48	.02	.02
76	59	59	55	0	.29
77	69	65	59	.25	1.69
78	70	66	60	.24	1.67
79	54	60	58	.60	.28
80	58	52	53	.69	.47
81	40	43	46	.21	.78
82	31	35	39	.46	1.64
83	25	28	32	.32	1.53
84	28	23	26	1.09	.15
85	21	20	21	.05	0
86	20	17	16	.53	1.00
87	9	14	13	1.79	1.23
88	10	11	10	.09	0
89	6	8	7	.50	.14
90	10	6	5	2.67	5.00
91	2	4	3	1.00	.33
92	3	2	2	.50	.50
93	2	1	1	1.00	1.00
94	1	1	1	0	0
95	0	0	1		1.00
Σ	675	675	676	15.77	23.75
Probability of fit P				.957	.694

Ave. $= \bar{x} = 78.846$ Ave. $= \beta_2 = 3.181$ Ave. $= \sigma_1 =$ .0943

$\sigma = 4.612$ $\sigma_z =$ .178 $\sigma_{\beta_2} =$ .189

$k =$ .521 $\sigma_\sigma =$ .126

* Phil. Mag., Vol. I, 1901, pp. 115-119.

high, meaning, of course, that 96 times out of 100 we may expect to find a system of deviations as large or larger than that actually found. The author finds, however, that the theoretical distribution

(column 4) based upon the assumption of the second approximation is also a very close fit to the observed frequencies, the probability of fit being in this case .69. As a result of these calculations shall we conclude that the distribution is composed of two normal components as indicated in Fig. 10, or shall we conclude that the distribution is homogeneous? In other words, do the skulls belong to two or to only

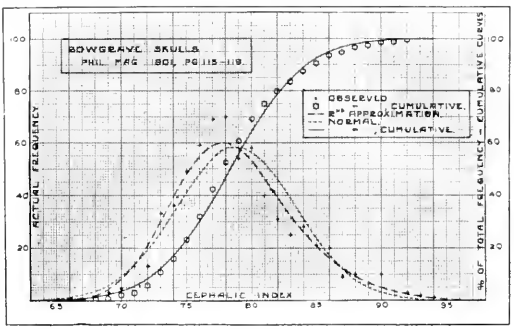


Fig. 10

one race? The measure given by the probability of fit is, of course, in favor of the first alternative. It is highly probable, however, that if we had been given the observed distribution without any discussion of what it meant we would have decided that it probably was consistent with the assumption of the random system of causes such as might underlie the second approximation.

In other words, if we had been given merely the above set of skull measurements, it is reasonable to suppose we might have concluded that the distribution was homogeneous. However, when our judgment is colored by the facts which cannot be presented in the array of observed frequencies we must conclude that it is highly probable that the observed data have arisen from a non-homogeneous population.

Statistical methods alone do not answer all of the questions that are raised in this problem nor do they answer them in many others. There is almost always room for judgment to enter.

Thus, analyzing a group of measurements of some characteristic of a large number of transmitters, it often becomes necessary to determine whether or not they can be subdivided into normal com-

ponents as in the above problem. In our case the subgroups correspond to different kinds of carbon. Here, as in the data given by Pearson, it often has been found necessary to base our final conclusion partly upon facts not revealed by the data themselves.

The integral curves corresponding to the normal and observed distributions are given in Fig. 10 in order to show that they do not

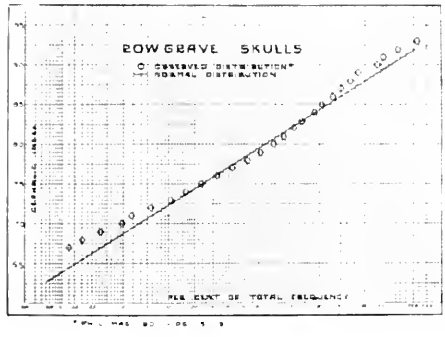


Fig. 11

serve to indicate the difference between the observed and theoretical distributions nearly as well as the actual frequency curves also given in this figure. Fig. 11 presents the result on probability paper. In this case the probability curves are as good as the frequency curves for showing the divergence between theory and observation. It will be recalled that this is not true for the similar curves given in Fig. 9.

Summary Statement of Suggested Method to be Followed in the Analysis of Engineering and Physical Data

We have briefly reviewed the different methods for determining the best theoretical distribution to represent observed data. The following four steps indicate the ordinary procedure:

1. Obtain the first four corrected moments.
2. Calculate the average, standard deviation, k and β_2, and their standard deviations.
3. Calculate the theoretical distribution of distributions warranted by the circumstances.

4. Apply one or more of the four methods of comparing the theoretical and observed distributions to determine which one is theoretically the best.[13]

An illustration of the method of applying this form of analysis to inspection data on transmitters is indicated in the schematic chart Fig. 12. The object of the inspection of apparatus in the process

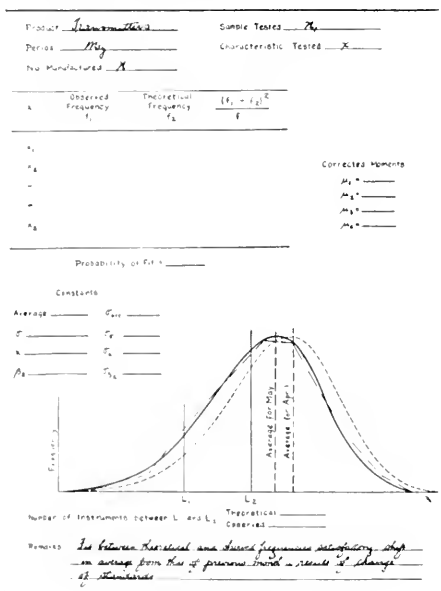


Fig. 12

of manufacture is obviously to determine the most probable law of distribution, and from this to determine whether or not there is any indication of a trend in the quality of the product. In the light of what has been said, it is obvious that a complete report of this character should contain the items called for in Fig. 12. The corrected

[13] If the observed distribution could not have arisen from a random system of causes, it may be advisable to attempt to transform it into an approximately random one, such as was done in connection with the data in Fig. 6.

moments and the factors, such as the average, standard deviation, k and β_2 should be given. These factors provide us with measures of the lack of symmetry, and can be used as pointed out in the previous sections of this paper. Recording this amount of data makes it possible for anyone interested, either to check the calculations of the theoretical frequencies and the conclusions derived therefrom, or to calculate a different theoretical distribution based upon fundamentally different hypotheses in a way such as has been illustrated already in the discussion of the distribution of measurements of the cephalic index, as given in Fig. 11.

In most instances, however, it is highly probable that the man who originally prepares the chart is charged with the responsibility of choosing the best distribution, and, therefore, the chief interest of those reading the report is centered upon the conclusions indicated therein. The graphical representation of the observed distribution by means of the histogram is hopeful. The comparison of this with the theoretical curve represented by a solid line shows qualitatively whether or not the product is changing. The probability of fit gives a quantitative measure of the degree of fit. The set of curves given in Fig. 12 is drawn to illustrate a condition which may sometimes happen when, for example, the standards used in the machines have been changed. This is only typical of the results which may be expected. Obviously, the form of such reports designed to meet specific conditions will vary. That presented above is only typical of one which has been found to be of value in presenting the analysis of the results of inspection of certain types of apparatus.

Some Advantages Derived from a Comparatively Complete Statistical Analysis

It has been pointed out that the value of either a physical or an engineering interpretation of data depends upon the success attained in deriving the best theoretical distribution. This is the equation which fits the observed points best, and which, if possible, can be interpreted physically. The previous discussion indicates the way in which different causal relationships tend to produce typical frequency distributions, and also the way in which statistical methods may be used in finding a theoretical distribution which yields a physical interpretation.

This point has been illustrated by several examples. It has been shown that by a proper choice of theoretical curve a very close approximation to an observed distribution can be obtained. This

TABLE VII—Frequency Distribution of σ Particles*

Number of Particles	Observed Frequency f	†Normal Law f_1	2nd Approximation f_2	‡Law of Small Numbers f_3	§Poisson Charlier f_4	$(p+q)^n$ f'	$f-f_1$ f_1	$(f-f_2)^2$ f_2	$f-f'$ f'	$(f-f_4)^2$ f_4	$\frac{(f-f_3)^2}{f_3}$
3	0	1	0	0	0		1 000				
−2	0	5	0	0	0		5 000				
1	0	23	8	0	0		23 000	8 000			
0	57	74	63	54	47	50	3 905	571	167	2 128	180
1	203	180	197	210	196	202	2 930	183	233	250	005
2	383	337	389	407	400	405	6 279	093	1 415	123	1 195
3	525	485	532	525	533	533	3 299	092	0	120	120
4	532	535	526	508	525	521	017	068	1 134	093	232
5	408	452	398	394	407	402	1 283	251	497	002	090
6	273	294	249	254	258	255	1 500	2 313	1 421	872	1 271
7	139	146	139	140	138	137	336	0	007	007	027
8	45	57	72	68	64	64	2 526	10 125	7 779	5 641	5 641
9	27	15	28	29	25	26	9 000	036	100	160	038
10	10	4	10	11	9	9	9 000	0	091	111	111
11	4	1	3	4	3	3	9 000	333	0	333	333
12	**0	0	1	1	1	1	∞	1 000	1 000	1 000	1 000
13	**1	0	0	0	0	0					
14	**1	0	0	0	0	0					
Σ	2608	2609	2615	2605	2606	2608	81 684	23 065	13 881	10 615	10 245
Probability of Fit							0000	0410	3086	5624	5946

$p = .046$
$q = .954$
$\bar{x} = pn = 3.87$
$\sigma = \sqrt{pqn} = 1.92$
$K = \frac{q-p}{\sqrt{pqn}} = .48$
$n = 84\ 174$

$\beta_3 = 3 + \frac{1-6pq}{pqn} = 3.51$

$\sigma_k = \sqrt{\frac{6}{2608}} = .0480$

$\sigma_{\beta_3} = \sqrt{\frac{24}{2608}} = .0959$

($N = 64$ used in computing $(p+q)^n$)

* Prof. Rutherford and H. Geiger—Phil. Mag., Oct., 1910.
† Normal calculated on basis of probability integral.
‡ Bateman, H., in appendix to original article.
§ Fisher, A.—loc. cit., p. 273.
** These observed frequencies are grouped together in computing P.

has already been indicated in Table III. To emphasize this point, however, let us consider once more the distribution of alpha particles given in Table I. These data together with various theoretical [1] distributions are given in Table VII.

Let us consider the data given in Table I by following the procedure of analysis outlined in the previous section. The factors k and β_2 when compared with their errors should indicate whether or not the distribution is normal. As shown in Table VII, k and β_2 differ from 0 and 3 respectively, by more than 3 times their respective standard deviations. As has already been pointed out, this is sufficient evidence to indicate that the distribution is not normal. In order to show, however, that if we follow the next step and calculate theoretical distributions based upon the assumption of the different laws; that is, in this case, normal, second approximation, and the law of small numbers, we are naturally led to the choice of the best distribution. This choice is materially influenced by the measure of the probability of fit as recorded in the table. The law of small numbers is obviously a very close approximation to the observed frequencies.

One of the obvious things to do in this problem, but one that has not been done previously, is to calculate the values of p, q and n, and from them the terms of the binomial expansion $2608(p+q)^n$. The probability of fit between the terms of this expansion and the observed frequencies is the highest given in the table. This increases the evidence that the distribution is random. It also does more. It serves to establish the facts that the probability p that an alpha particle will strike the screen is .016, and that the maximum number of alpha particles which may ever be expected to strike the screen is of the order of magnitude of 81. Granted then that we can always find the most probable theoretical frequency distribution, let us consider next the influence that the result may have in our determination of the most probable value, the number of observations between any two limits and the casual relationships governing the distribution.

Let us consider first the dependence of the most probable value upon the type of distribution. In our present work in the study of carbon the resultant distributions have been in most instances either random or such that through a proper transformation they could be reduced to such. For any distribution consistent with the second approxima-

[1] The source of all distributions previously calculated are indicated. The Poisson-Charlier series is similar to the Gram-Charlier series, except that the law of small numbers is the generating function. It serves as an admirable method of graduating certain classes of skew distribution as illustrated by this example and by that given in Table III.

tion the most probable value is at a distance $-\frac{k\sigma}{2}$ from the arithmetic mean. Many distributions have been found for which k lies between .5 and unity, and, therefore, this difference is from $\frac{1}{4}$ to $\frac{1}{2}$ of the standard deviation. Thus, the efficiencies of certain standard types of transmitters are found to conform to such a law, and the difference between the modal and average values is of the order of magnitude of 0.4 mile.

Obviously the geometric mean of the sound intensities (Fig. 6) and not their arithmetic mean is the most probable. The difference between the two is quite large. The difference between the arithmetic mean and the modal value for groups of data such as given in Fig. 1, Tables II and VI are quite large. To use again the illustration of the alpha particles the observed most probable number is 4; whereas, the observed average [45] is 3.87. Judging from the best theoretical distribution the most probable number of alpha particles is 3. Choosing the number 3 it is seen that either of the other two numbers differ from this by approximately $\frac{1}{2}$ the standard deviation. Such results are, however, not confined to the work of the present investigation nor to the examples previously cited as is evidenced by the data given in the last column of Table VIII.

TABLE VIII

N = Number of Observations	Source of Data	Percentage Within $\bar{X} \pm \sigma$	Percentage Within $\bar{X} \pm 2\sigma$	Percentage Within $\bar{X} \pm 3\sigma$	Average — Modal / σ
1000	E *54	66.6	97.2	99.6	.803
251	E 66	78.1	94.8	97.6	1.042
9154	E 10	67.7	95.5	99.6	.031
2162	E 79	70.1	95.1	99.3	−.311
368	E 84	73.4	94.6	97.0	.422
675	Table VI	68.7	94.1	99.6	.247
	Normal Law	64.26	95.44	99.73	0

* Elderton "Frequency Curves and Correlation," published by C. & E. Layton, London, 1906.

We should not leave this phase of the discussion, however, without pointing out that in a large number of purely physical experiments a sufficient number of observations has not been taken to make it possible to choose the best theoretical distribution. In general more than

[45] Of course, such an average has no significance, except for a continuous distribution.

100 observations are required. Thus, in Prof. Millikan's [46] determination of the electron charge e only 58 observations were made. The values of σ, k, and β_2 for this distribution are .428 units, —.496 and 2.358. Even though the observed distribution is consistent with a normal system of causes, values of k and β_2 may be expected to occur which differ from 0 and 3 respectively, as much as these observed values do. In this case even if k is real and not a result of random sampling, the correction to be added to the average in order to obtain the most probable value is insignificantly small.

Next let us consider the problem of determining the number of observations between any two limits. The physicist is ordinarily concerned with the probable error: that is, the error such that $\frac{1}{2}$ of the observations lie within the range $X \pm$ probable error. Its magnitude for the normal distribution is $.6745\sigma$, and the errors are distributed symmetrically on either side of the average. It is interesting to note that the magnitude of the probable error is also $.6745\sigma$ for the second approximation, but that the errors are not distributed symmetrically on either side of the average.

Another important pair of limits is that including the majority of the observations. For the normal law 99.73% of the observations are included within the range $X \pm 3\sigma$ which, therefore, is often called the range. Not a single example has been found, however, of a distribution for which the observed number of observations within this range is less than 95% even though the distribution is decidedly skew. In fact it is seldom less than 98%. If, however, we have a case such as that represented in Table II where groups of observations have been taken in what is technically known as different universes, and then averaged together, the average result is not the most probable, and the standard deviation of the average is not inversely proportional to the square root of the number of observations. Since this point is of considerable importance, it is perhaps well to state it in a slightly different way. Thus, let us assume that we have a thousand samples of granular carbon which possess inherent microphonic efficiencies differing by comparatively large magnitude. Transmitters assembled from any one of the groups of carbon cover a range of efficiencies. If we choose a sample of 10,000 instruments, 5,000 from each of two lots of carbon which do not possess the same inherent efficiency, we cannot expect, for reasons already pointed out, that the observed distribution will be normal. The average of these observations will not in general be the most probable value, and the standard deviation of the average will not be equal to the

[46] Millikan, R. A.—The Electron—*University of Chicago Press.*

observed standard deviation divided by the square root of the number of observations, in this case 10,000.

We have already seen, however, that it is possible to detect such errors of sampling, since in general the distribution cannot be fitted by the second approximation or Gram-Charlier series. If the theoretical distribution is either normal, second approximation, or the law of small numbers, the number of observations to be expected between any two limits can be readily determined from the tables. Experience has shown that in every instance where it has been possible to represent the observed distribution in any of these three ways, the data obtained in future samplings have always been consistent with the results to be expected from the theory underlying these three laws. It will be of interest to note the data given in columns 3, 4, and 5 of Table VIII and to compare the theoretical percentages (last row) for the different limits with those observed.

In closing it is of interest to point out further the significance of some of the results discussed in this paper in connection with the inspection of equipment. Here we must decide upon a magnitude of the sample to be measured in order to determine the true percentage of defective instruments in the product. If p is the percentage defective, and q that not defective, then the standard deviation about the average number found in a sample of n chosen from N instruments is

$$\sigma^2 = pqn\left(1 - \frac{n}{N}\right).$$

In practice, however, we never know the true value of p unless we measure all of the apparatus, and this is impractical. In our calculations we must therefore use some corrected value. We find, though, that the average value of p is in most instances the one that must be used. Assuming that we choose a value of p, the distribution of defectives in N' samples of n in number will be represented by the distribution of $N'(p+q)^n$. If one of the samples is found to contain a percentage of defectives, which is inconsistent, that is, which is highly improbable as determined from the distribution of $N'(p+q)^n$, it indicates that the product is changing.

If, however, we take into account the effect of the size of the first sample in respect to the second as indicated by Pearson,[47] we see that the distribution of N' samples may be different from that given by the binomial expansion. In accordance with this theory, if in a first sample of 100, 10% of the sample is found to possess a given attribute,

[47] Pearson, K. Loc. cit. Foot note 30.

the distribution of the percentages to be expected in 1,000 such samples is indicated by the last column of frequencies in Table III. In order to show graphically how this distribution differs from that

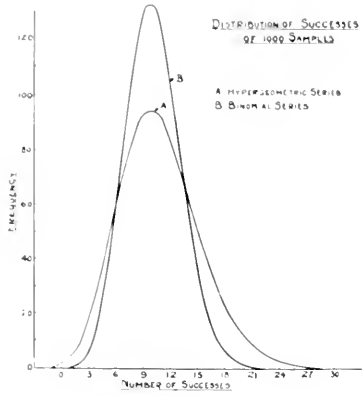


Fig. 13

corresponding to the binomial expansion these two sets of frequencies are reproduced in Fig. 13. The difference between them is a striking illustration of the significance of the size of the samples used in connection with the inspection of equipment, providing we accept Pearson's results.

Deviation of Random Samples from Average Conditions and Significance to Traffic Men

By E. C. MOLINA and R. P. CROWELL

THE traffic executive deals with questions which lead him into the consideration of problems of widely differing natures. At almost every turn he is confronted by the fact that his decisions and programs in relation to these different phases of the work must be based on records which are seldom continuous and in most cases are merely "samples." These sample records are assumed to measure the characteristics of the entire volume of facts or data of which they are taken to be representative. In the use and analysis of these records there are a number of perplexing questions which come to his mind if he allows himself the luxury of a little theoretical speculation.

Practically all of his information regarding the efficiency with which his office is run and on which he must base his plans for continued efficiency is obtained from the peg counts. These peg counts are records of the number of calls handled and are taken on two or three days out of each month. At the same time that the calls are counted, the number of employee hours used in the handling of the traffic is counted. The results of these peg counts are used to represent the performance of that office for the month. When the inquiring traffic man meditates a little on the subject of these peg counts he soon begins to wonder how nearly representative they are of his every day performance. He can—and sometimes does—think up a number of things which will explain any poor results which show up.

One of the means taken to insure the accuracy of the peg count is to observe the counting of 25 to 50 calls each by as many of the operators as possible, with the idea of determining how accurately the operators count. In this way from 1,500 to 3,000 observations are made on the accuracy of the operators' counting, in a period of two or three days. The traffic man occasionally questions whether he can rely on the results of this comparatively small number of checking observations to give him an indication of the accuracy of the count as a whole.

In order that comparisons may be made of the performance of different offices and the cost of handling different kinds of calls, it is the practice to translate all the work done into terms of traffic units (representing the relation of the labor value of the different operations to a fixed value arbitrarily selected). In order to do this, at longer intervals than the regular peg counts, the traffic is counted in

more detail. From certain classifications and subdivisions of these supplementary counts, coefficients or equating factors are developed which are applied to the regular counts to develop units. The speculative traffic man ponders over these and wonders how representative the supplementary counts are of the every day distribution of traffic.

This speculation leads him also to question the labor values which have been assigned to the different operations and which have been furnished him for the purpose of equating his traffic. He knows that because of the impossibility of making continuous stop watch observations on his operators, he has to accept the results of such observations made on a considerable number of calls handled in a similar manner at some time in the past and probably in some other place, as being representative of the work involved in handling those types of calls at the present time in his office.

After thus puzzling himself over peg counts and similar records, the traffic man may turn his attention to some of the service problems and begins to scrutinize with considerable skepticism the records which are maintained of this feature of his work. Among the most valuable records of the way in which the service at his office is being handled, are the records developed as a part of the central office instruction routines. These are observations taken on ten calls handled by each of the operators on the force, periodically. He looks over the latest detail sheets and observes that the results of these tests on two particular operators show that the one he considered a very careful and methodical girl has made a high proportion of mistakes while the operator whom he thinks is the more careless shows an absolutely perfect test. Because of his other knowledge he suspects these records and decides to check them up by examining the summaries of similar tests taken for some months past. These summaries show figures which bear out his original estimate of the ability of the two operators, which relieves his mind but leaves him still puzzled as to why the averaging of a series of figures which are not representative, makes the summary more nearly representative.

There is another set of figures which the traffic man consults in connection with the quality of the service and which causes him a good deal of worry. These are the figures obtained from central office speed of answer tests, tests of the speed of answer to recall signals, etc. The speed of answer tests, for example, are made by an employee in the central office who causes signals to appear and with a stop watch determines how long it takes the operators to answer each signal. The signals used in making these tests are distributed in all parts of the switchboard and the number of tests made in each

hour is roughly proportional to the amount of traffic handled. The results of these tests are summarized in such a manner as to show the percentage of tests which are not answered within 5, 10 and 20 seconds. The traffic man who gives this matter thought, is concerned to know how much reliance he can place on the results of these tests as being representative of the percentage of slow answers applying to all the calls handled in the office.

The speculative traffic man by this time is in a frame of mind which either leads him to doubt all figures or to feel that there must be something in the figures which he cannot explain but which makes certain of them quite representative, although there are certain others about which he does not feel the same way. He is sure that some of them are representative because decisions and programs based on them produce the results desired. He is also sure that some of them are not representative because they imply things which he knows are not so, as a result of observation. Just how far he can rely upon the figures which he is using, and where to draw the line is a question which only long experience or an understanding of the reasons which lie behind the taking of these records can solve. It will probably be of interest to discuss, from the purely theoretical angle, certain simple traffic data with the idea of noticing how the application of a certain mathematical procedure can aid in drawing accurate conclusions from them.

The type of traffic problem which will be considered may be stated as follows:

A group of 50,000 calls originated in an exchange area. An unknown number of them were delayed more than 10 seconds. Observations were made on 300 of the calls and of these 9, or 3 per cent., were delayed more than 10 seconds. With this information is it a safe bet that the unknown percentage for the entire 50,000 calls is below 5? Or better yet, are we justified in betting 99 in 100 that the unknown percentage for the 50,000 calls is below 5? Or again, may we bet 8 in 10 that the unknown percentage is between 0.5 and 5? It is taken for granted that the observer is justified in believing that the calls under consideration fulfill the conditions of random sampling such as that each call is independent of every other call, or that an appreciable number of the calls is not due to the occurrence of some unusual event, the opening of the first game of the world series, for example.

Assuming that the reader is unfamiliar with the theory of probability, a digression becomes necessary and in order that he may enter into the spirit of the theory the reader is requested to forget for the

present the telephone problem. Of course, only a bird's-eye view of the theory will be given here. Several lacunæ will be encountered, the filling in of any one of them would call for a volume of not very small dimensions.

Introduction to the Theory of "A Posteriori" Probability

The problem to be dealt with belongs to the class of problems which gave rise to that branch of the Theory of Probability which is known as "A Posteriori Probability" or "Probability of Causes." It is frequently referred to as the Theory of Sampling.

To bring out certain of the ideas involved it will be helpful to consider what may appear as a very extreme example from the traffic man's point of view, but which is nevertheless typical of the type of problem in which a consideration of a posteriori probability enters. We are told that at a student gathering a particular young man won 7 out of 15 times. Our informant refuses to divulge what is going on at the gathering. What probabilities should we assign to the following hypotheses?

1. He threw heads 7 times out of 15 throws with a coin.
2. He threw 7 aces out of 15 throws with a 6 face die.
3. He won on points 7 rounds in a fifteen round bout.
4. The aggregate of all other hypotheses.

A little careful consideration will make it clear that with reference to each hypothesis (or aggregate of hypotheses) two essential questions must be answered before we can determine the a posteriori probability. Consider the six face die hypothesis; we must know:

1st—What is the relative frequency or probability with which gambling with a 6 face die is indulged in at student gatherings?

2nd—Given a six face die, what is the probability of throwing an ace 7 times in 15 throws?

Quoting Mr. Arne Fisher[1] we may restate these two questions as follows:

1st—What is the a priori *existence* probability in favor of the 6 face die hypothesis?

2nd—What is the *productive* probability for the observed event given by the hypothesis of a 6 face die?

[1] Arne Fisher—The Mathematical Theory of Probabilities—2nd Edition—Art. 41.

In most problems of this type the determination of the *productive* probability for each hypothesis is a question of pure mathematics. But when we proceed to evaluate the a priori *existence* probability for each hypothesis or cause, common sense and guessing must frequently be resorted to. The history of the applications of a posteriori probability is so full of paradoxes resulting from appeals to common sense that to some high authorities the whole theory is a fallacy. Prof. George Chrystal [2] closes a severe attack on Laplace's Theorie Analytique with the statement—"The indiscretions of great men should be quietly allowed to be forgotten." Nevertheless, the writers will assume the Laplacian view of the subject, especially as it has been defended by such authorities as Karl Pearson and E. T. Whittaker.

The above typical problem has been introduced because its mere statement leads us immediately to the conceptions of existence and productive probabilities with reference to different possible hypotheses. But, it is not our intention to bring any notoriety on the young man by answering the questions raised. Moreover, the hypotheses made, differ qualitatively, whereas, our telephone problem involves various hypotheses which differ only quantitatively. We, therefore, proceed to another typical problem, a solution of which will give us at once the solution of the telephone problem.

A bag contains 1,000 balls; an unknown number of these are white and the rest not white. Of 100 balls drawn 7 are found to be white. What light does this information throw on the value of the unknown number of white balls? What is the probability that there are 70 white? Is it a safe bet that the number of white balls lies between 60 and 80?

Two cases of this problem may be considered:

Case 1. After a ball is drawn it is replaced and the bag is shaken thoroughly before the next drawing is made.

Case 2. A drawn ball is not replaced before another ball is drawn.

These two cases become essentially identical if the total number of balls in the bag is very large compared with the number drawn.[3] In the following discussion Case 1 is assumed.

The information at hand is that 100 drawings resulted in 7 whites. Obviously the bag contains at least one white, but we are free to choose between 999 possible hypotheses.

[2] Transactions of the Acturial Society of Edinburgh—Vol. II, No. 13—On Some Fundamental Principles in the Theory of Probabilities.

[3] For the application to practice herein contemplated it is thought that the number of balls in the bag should be at least ten times the number drawn.

1—The bag contains 1 white and 999 not white.
2—The bag contains 2 white and 998 not white.
3—The bag contains 3 white and 997 not white.
. . . . . .
K—The bag contains K white and (1,000-K) not white.
. . . . . .
997—The bag contains 997 white and 3 not white.
998—The bag contains 998 white and 2 not white.
999—The bag contains 999 white and 1 not white.

Let $H(K)$ be the existence probability for the K'th hypothesis. By "existence probability" is meant the likelihood that the bag contains exactly K white balls when the circumstances of the drawing, but not the actual results of the drawing, are fully taken into account. Its exact value may often be in doubt either because we do not have complete knowledge of the circumstances preceding the drawing or because we are not able to deduce its exact value from this knowledge. It is obvious, however, that there must be some such value and we must, therefore, introduce a symbol to represent it.

Let $B(7,100,K)$ = productive probability for the K'th hypothesis; by this is meant the probability of obtaining the observed event (7 white in 100 drawings) if the bag contains K white balls and 1,000-K that are not white.

Then the a posteriori probability in favor of the K'th hypothesis (meaning thereby the probability in favor of the K'th hypothesis *after* the 7 white balls were drawn) is [4]

$$P_k = \frac{H(K)B(7,100,K)}{\sum_{s=1}^{s=999} H(S)B(7,100,S)} \quad (1)$$

Now to say that the bag with a total of 1,000 balls contains K white balls is equivalent to saying that the *ratio* of white to total balls is

$$p_k = K \div 1000$$

and that the *ratio* of not white to total balls is

$$q_k = 1 - p_k = (1000 - K) \div 1000.$$

[4] This is the celebrated Laplacian generalization of Bayes' formula. No attempt to demonstrate it will be made here. The subject is dealt with at length by Laplace in the Théorie Analytique des Probabilités and by Poisson in the Recherches Sur La Probabilité des Jugements. A beautiful and relatively short demonstration is given by Poincaré in his Calcul des Probabilités.

We may, therefore, rewrite (1) as follows:

$$P_k = \frac{\Pi''(p_k)B'(7,100,p_k)}{\sum_{s=1}^{s=999}\Pi''(p_s)B'(7,100,p_s)}, \tag{2}$$

where Π'', B' are the forms assumed by the functions W, B, respectively, when the ratio p_k is used instead of the number K.

The interpretation of the terms of the expansion of the binomial $(p+q)^{100}$ tells us that

$$B'(7,100,p) = \binom{100}{7} p^7(1-p)^{93} = \binom{100}{7} p^7 q^{93}$$

where $\binom{100}{7}$ is a symbol for the number of combinations of 100 things 7 at a time.

Substituting in (2) and canceling from numerator and denominator the common factor $\binom{100}{7}$ gives

$$P_k = \frac{\Pi''(p_k)p_k^7(1-p_k)^{93}}{\sum_{1}^{999}\Pi''(p_s)p_s^7(1-p_s)^{93}}. \tag{3}$$

From (3) we obtain for the a posteriori probability that the ratio of white balls does not exceed K_2 1,000,

$$P(K \geq K_2) = \sum_{1}^{K_2} P_k.$$

Likewise, the a posteriori probability that the ratio is not less than K_1/1,000 is

$$P(K \leq K_1) = \sum_{K_1}^{999} P_k.$$

Finally, the a posteriori probability that the ratio is not less than K_1/1,000 or greater than K_2 1,000 is

$$P(K_1 \leq K \leq K_2) = \sum_{K_1}^{K_2} P_k = \frac{\sum_{K_1}^{K_2}\Pi''(p_s)p_s^7(1-p_s)^{93}}{\sum_{1}^{999}\Pi''(p_s)p_s^7(1-p_s)^{93}}. \tag{4}$$

Solution of the Telephone Problem

Obviously the telephone problem is analogous to the problem of the bag containing an unknown ratio of white balls. The corresponding elements in the two problems may be tabulated as follows:

1st—1,000 balls in bag versus 50,000 calls originated.
2nd—100 balls drawn versus 300 calls observed.
3rd—7 white balls drawn versus 9 calls delayed more than 10 seconds (*i.e.*, defective with reference to a particular characteristic).
4th—To the 999 possible hypotheses with reference to the unknown per cent. of white balls correspond 49,999 possible hypotheses with reference to the unknown per cent. of calls delayed more than 10 seconds.

The problems differ in that a ball drawn from the bag is returned before another drawing is made, whereas an observed call is comparable to a ball being drawn and not returned. With the numbers involved, however, the discrepancy may be ignored.

A formula of the same form as (4) will, therefore, give the answer to our question. We may, however, substitute definite integrals in place of the finite summations since the difference between any two consecutive possible values for the unknown ratio is very small. The integrals together with some desirable transformations of them will be found in the appendix to this article. We will mention here, however, that the transformations made involve an arbitrary assumption as to how the a priori *existence* probability for the different hypotheses varies. As stated above in connection with Prof. Chrystal's views, this is the phase of the subject which lends itself to considerable difference of opinion. The reader who contemplates using the curves embodied in this article should read the appendix with special reference to the assumptions made.

The attached curves Fig. 1 show graphically the conclusions to be drawn from the mathematical analysis. A glance at the right hand end of the curves will show that they are associated in pairs. The upper curve of a pair slopes downward from left to right while its mate slopes upward.

Consider the pair of curves marked .03. For the abscissa 300 they give as ordinates the values .0625 and .014. The interpretation of these figures is as follows: if 300 observations gave 3 per cent. of calls delayed then we may bet

1st—99 in 100 that the unknown percentage of calls delayed is *not greater* than 6.25.

2nd—99 in 100 that it is *not less* than 1.4 per cent.
3rd—98 in 100 that it lies between 1.4 per cent. and 6.25 per cent.

Likewise, considering the curves marked .06 if 1,000 observations gave 6 per cent. of calls delayed, then we may bet

1st—99 in 100 that the unknown percentage of calls delayed is not greater than 8.05.
2nd—99 in 100 that it is not less than 4.4 per cent.
3rd—98 in 100 that it lies between 4.4 per cent. and 8.05 per cent.

It is obvious from the shape of the curves that a few hundred observations do not give more than a vague idea as to the unknown per cent. of calls delayed. On the other hand, the gain in accuracy obtained by making more than 10,000 observations would hardly justify the expense involved. The number of observations which safety requires in any particular problem must be determined by the conditions of the problem itself. If we are willing to take a chance of 9 in 10 or 8 in 10 instead of 99 in 100 or 98 in 100, respectively, the curves of Fig. 2 will give us an idea of the range within which the unknown percentage of defectives lies.

APPENDIX

Case No. 1—Infinite Source of Samples

An inspection of n samples has given c defectives. The observed frequency is then c/n. Let p be the unknown true frequency and p_1 the frequency of delayed calls which has been arbitrarily chosen as being the maximum permissible.

The a posteriori probability that $p \gtrless p_1$ is

$$P = \frac{\int_o^{p_1} H'(x) x^c (1-x)^{n-c} dx}{\int_o^1 H'(x) x^c (1-x)^{n-c} dx}, \tag{1}$$

where $H'(x)$ is the a priori existence probability that $p = x$. This formula is unmanageable if the form of $H'(x)$ is unknown.

Assume first that $H'(x)$ is a constant b for $o < x < g$, where $g > p_1$. Then

$$P = \frac{\int_o^{p_1} x^c (1-x)^{n-c} dx}{\int_o^g x^c (1-x)^{n-c} dx + \int_g^1 \frac{H'(x)}{b} x^c (1-x)^{n-c} dx}. \tag{2}$$

Now assume that

$$\int_{\epsilon}^{1} \frac{\Pi'(x)}{b} x^{c}(1-x)^{n-c}dx,$$

is negligible compared with

$$\int_{0}^{\epsilon} x^{c}(1-x)^{n-c}dx,$$

and also assume that g, c and $(n-c)$ are such that approximately

$$\int_{0}^{\epsilon} x^{c}(1-x)^{n-c}dx = \int_{0}^{1} x^{c}(1-x)^{n-c}dx.$$

Then, finally,

$$P = \frac{\int_{0}^{p_1} x^{c}(1-x)^{n-c}dx}{\int_{0}^{1} x^{c}(1-x)^{n-c}dx} = \frac{(n+1)!}{c!(n-c)!}\int_{0}^{p_1} x^{c}(1-x)^{n-c}dx, \tag{3}$$

This well known formula might have been obtained by assuming *ab initio* that $\Pi'(x)$ is independent of x. It should be particularly noted that this independence is not identical with the assumptions made above. In the applications which are here contemplated the values of p_1, c and n are such that g need be but a small fraction of the range o to 1.

In the "Théorie Analytique" Laplace transforms (3) so that it can be evaluated in terms of the Laplace-Bernoulli integral

$$\frac{2}{\sqrt{\pi}}\int_{0}^{k} e^{-t^2}dt,$$

where k is a function of p_1, c and n. This transformation is most valuable when p_1 is in the neighborhood of 1/2. For small values of p_1 the transformation which converts the binomial expansion to Poisson's exponential binomial limit is more appropriate and gives, writing $(n\ p_1) = a_1$,

$$P = \frac{1}{c!}\int_{0}^{a_1} y^{c} e^{-y}dy = P(c+1, a_1). \tag{4}$$

Photomicrography and Technical Microscopy in Its Application to Telephone Apparatus

By FRANCIS F. LUCAS

NOTE. The following paper may be considered as introductory to the subject of photomicrography. Doubtless everyone is casually familiar with photomicrographs of the crystalline structure of various metals. The application of this branch of the optical art to the study of metals is very important in the design and manufacture of telephone apparatus but its importance in telephony is more far-reaching than in the study of metals alone. Various of these applications are suggested by the illustrations reproduced in the Appendix of this article. *Editor.*

INTRODUCTION

BY photomicrography is meant the adaptation of photography to microscopy, or the art of photographing a magnified image. The scope of the art embraces the reproduction of images ranging from natural size up to magnifications of several thousand times, the degree of magnification being expressed in terms of diameters. It will be seen that the image is not always magnified but in some instances may be at a 1:1 ratio or when large subjects are being photographed, at an actual reduction in size. Such low-power work is often spoken of as gross photography but so far as the equipment and technique of treatment is concerned it is low-power work.

Low-power work may be considered as treating with magnitudes from 1 to about 30 diameters. Medium-power work deals with magnifications from about 30 to about 500 diameters, and high-power work extends from 500 diameters upward. The limit of useful magnification is a much disputed question. It is sometimes contended that 1,500 diameters represents about all that is worth while, but the fact that very few pictures are published which exceed 1,500 diameters in magnification would lead to the conclusion that either the limit is from 1,000 to 1,500 or else the art has not been developed to the state where substantial gains result by going higher. This matter will be considered at greater length below.

GENERAL DISCUSSION OF APPARATUS

The reason that photomicrography is grouped under three classifications according to magnification, is because the apparatus used in each case is quite different and because the preparation of the subject and its treatment also differ. In fact for low-power work the microscope often may be dispensed with entirely, the lens being secured directly to the camera; in other cases, the microscope serves only as a

convenient support for a lens. In the treatment of most transparent mounts an illuminating device termed a substage condenser is necessary, the microscope then forms a very necessary adjunct to low-power work.

Medium-power work always requires the use of a microscope, and because rigidity in mounting and accuracy in adjustment are very necessary to correct rendering of the image, some sort of a stand is provided on which the microscope and a suitable illuminating train are mounted. Usually this stand takes the form of a narrow wooden or metal table supported by substantial metal legs. The table carries an optical bench which in practice is a metal bar or rail of special and rugged construction upon which the optical parts, the illuminant and the camera are mounted and are capable of adjustment so that they may be aligned optically. The description necessarily, meets generalized conditions. There is, however, a great similarity in the product of different makes of equipment and they all follow the same conventional lines, improvements in one make quite often being met by similar changes on the part of other makers.

There is no very well defined line between medium-power and high-power apparatus so far as the stands are concerned, but when it comes to real precision apparatus the choice in equipment is limited to possibly two or three makes. The difference is to be found in the quality of the optical parts and in the general stability of the assembly. A skilled technician may produce remarkable medium-power results with quite ordinary apparatus but no amount of training and skill can make good in high-power work for the actual shortcomings of an objective. Given a really good objective the skilled operator may use an inferior type of stand and secure very fine results, but he will be working under a considerable handicap and his work will not be consistently good because lack of the right sort of apparatus is apt to introduce variations in illumination, focusing, or adjustments which will prove ruinous to good definition.

Thus far consideration has been given to apparatus capable of yielding a magnified image of some tangible sort of a specimen, but there is an entirely different form of microscopic equipment which reveals the presence of particles beyond reach by all other known means of microscopic vision; reference is made to the ultra-microscope. This instrument is not ordinarily provided with photographic apparatus although with certain classes of work and under favorable conditions it is possible to reproduce the image photographically. Both liquids and solids may be studied by this means but in each case the specimen must be capable of transmitting light.

The Compound Microscope

It is obvious that a complete technical discussion of the instrument and equipment used in photomicrography is not within the scope of this paper nor would it be of interest to many readers. In order to appreciate the possibilities of technical microscopy as an aid in the

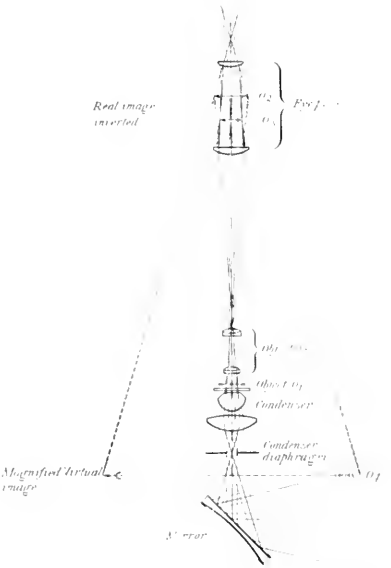


Fig. 1—Optical Diagram of the Compound Microscope

solution of definite engineering problems relating to telephone apparatus it is necessary, however, to consider more in detail the equipment used.

The optical system of the compound microscope is shown diagrammatically in Fig. 1, and in Figs. 2 and 3 are pictured two modern representative research type microscopes. In the diagram three parallel pencils of light are shown reflected upward into the condenser which by proper focusing is caused to illuminate a transparent object (suitably prepared and mounted as described later) placed in position on the microscope stage. As shown the objective would form

an inverted real image of the object O_1 at O_2 but the rays are intercepted by the lower lens of the eyepiece before the real image is formed. The lower eyepiece lens in combination with the upper

Fig. 2—Research type of microscope by Zeiss. Large barrel for photo-micrography; revolving mechanical stage, and sliding objective changers.

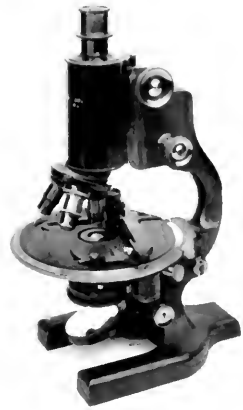


Fig. 3—Research type of microscope by Spencer Lens Co. Large barrel for photo-micrography; a large revolving stage with graduated circle, and a removable mechanical stage.

eyepiece lens forms a magnified virtual image O_3 of the real image O_2. There are two magnifications of the object and the resulting final magnification is the product of the magnifying powers of the objective and the eyepiece.

It should be noted that the objective produces an enlarged image of the object and that the eyepiece further magnifies this image; from this it is evident that if detail is lacking or if the image is not a good likeness of the object, the eyepiece will not make up for the shortcomings of the objective. The objective, then, becomes perhaps the most important part of the whole outfit. No one objective will serve for all purposes because of the limited range throughout which each particular objective is most useful; hence it is necessary to have a whole battery available so that the objective may be selected to suit the requirements of the work.

Objectives are divided into four general classes: achromatic, semi-apochromatic, apochromatic and monochromatic for use with ultra-violet light. These objectives do not consist of single lenses but are composed of two or more lenses very accurately centered and permanently mounted in a metal holder. The component parts of the lens system are chosen with regard to their ability to correct or compensate

Fig. 4—A battery of low-power lenses. These lenses are used without eyepieces. Each lens is equipped with a diaphragm for stopping the aperture.

for certain errors which are always characteristic of a simple lens. The value of an objective depends on the degree to which these imperfections have been overcome.

The difference in quality between the first three classes of objectives is primarily a matter of correction for chromatic and spherical aberrations. Chromatic aberration is the inability of a lens to focus sharply at the same point the different colors which go to make up the incident light and the inability to bring two rays of incident light of the same color to the same focus is termed spherical aberration.

The achromatic objectives have the chief optical defects corrected in a sufficient degree for the physiologically most effective rays (yellow-green) of the visible spectrum, while in the case of the apochromatic objectives the correction of the image defects extends approximately evenly over the entire range of the visible spectrum from the red to the violet regions.

In the achromatic lenses the fusion of the chromatic rays becomes less and less complete for rays belonging to the extremes of the visible spectrum under the ordinary conditions of illumination with white light, and this imperfection becomes more apparent when highly magnifying eyepieces are used. There are also residual imperfections in the fusion of the rays so that the colors of objects are not rendered with absolute precision in their finer shades. In the apochromatic objectives the fusion of the rays is so perfect that they may be used in conjunction with high-power eyepieces, and because of this perfect fusion the natural colors of the object are rendered with

great precision. The semi-apochromatic objectives contain fluorite elements and these objectives occupy a position in quality intermediate between the achromatic and apochromatic types.

Objectives are classified and listed according to their optical characteristics such as primary magnification, numerical aperture and focal length and as to whether they belong to the "dry" or the "im-

Fig. 5.—A battery of medium and high-power objectives and eyepieces

mersion" series. The term "dry" signifies that the objective when properly used is separated from the specimen by a stratum of air. In the case of the immersion objectives some one fluid for which medium the objective has been computed, such as water, glycerine, cedarwood oil, etc., is used to connect the front lens of the objective with the specimen. The fundamental difference between the dry and the immersion objectives is one of resolution, where by resolution

is meant the ability to see separate and distinct lines as individual units when these lines are spaced very close together. Resolving power or the number of lines per inch resolved is expressed numerically by the equation

$$N = \frac{2\ N.A.}{\lambda},$$

in which N is the number of lines per inch, $N.A.$ is the numerical aperture (defined below) and λ is the wave-length in inches. An objective of high resolving power when correctly used will resolve

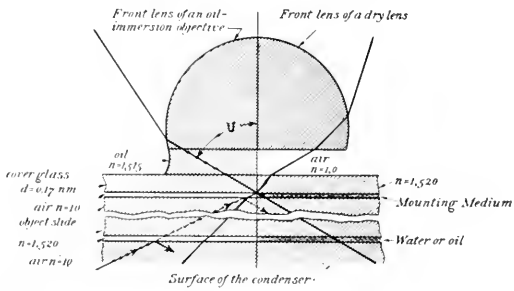


Fig. 6—Diagram illustrating numerical aperture and the superior light gathering powers of an oil immersion objective.

lines spaced 100,000 to the inch, whereas an objective of inferior resolving power under the same condition will not be able to distinguish the lines as distinct units.

As will be seen from Fig. 6, an immersion lens has greater light gathering power than a dry lens of corresponding focal length. This light gathering power is expressed as numerical aperture which term in reality supplies a measure of all of the essential qualities of the objective. The magnitude of the numerical aperture is expressed by the equation

$$N.A. = n \sin U,$$

n being the refractive index of the medium contained between the cover-glass and the front lens of the objective, and U the semi-aperture angle of the system.

For a given magnification and under comparable conditions the resolving power is directly proportional to the numerical aperture. The brightness of the image is proportional to the square of the

numerical aperture. As the numerical aperture increases the depth of penetration (*i.e.*, the power of the objective to resolve detail simultaneously at different depths or distances from the objective), and the flatness of the field both decrease, but usually when high resolution is desired flatness of field and penetration are not of great concern. The value of the numerical aperture varies from about 0.10 in the very low-power achromatic objectives to 1.40 for the oil immersion apochromats.

The eyepieces for use with the achromatic objectives are generally of the Huygens type but those for use with apochromats are termed compensating because of certain corrective measures which they apply to the behavior of this type objective. High-power achromatic objectives and the semi-apochromats may also be used to advantage with the compensating eyepieces. The magnifying power of eyepieces ranges from about 4 times to 20 times although another class termed orthoscopic eyepieces may be procured with a magnifying power of 28. These latter eyepieces are generally used with low-power objectives only. A special type of eyepiece known as a projection eyepiece of low magnifying power is used for certain classes of work when photographing with a long bellows extension. These eyepieces have correction collars which must be set to correspond with the bellows extension used.

Illumination

The color of the light used and the illumination of the specimen play a most important part in photomicrography and the behavior of the finest objective will appear very ordinary unless critical illumination of the specimen is attained. The illuminant is usually some form of arc lamp or metal filament, gas-filled lamp. Both types have their advantages and while many statements may be found derogatory to the use of arc lamps, as a real source of light, the author has found that a smoothly operating automatic arc lamp equipped with suitable carbons is capable of yielding results of the highest order. What is needed especially for medium and high-power work is a point source of light (or approximately so) of great brilliancy; capable of being smoothly and uniformly controlled so that the luminous end of the positive carbon will not fluctuate backward and forward within wide limits. Most automatic arc lamps are designed for a certain direct current value, usually about five amperes and unless the current rating is closely adhered to in practice the operation is apt to be irregular. Sputtering and irregular feeding

of the carbons are due to lack of proper adjustment of apparatus or of current and voltage or to the use of an unsatisfactory grade of carbons. It may be of interest to know that the type of automatic arc lamps used in the Bell System Laboratory are so steady and uniform in their operation that they occasion no concern except for the usual maintenance. Since very small diameter carbons must be used to approximate the point source of light condition, these lamps will operate continuously with one set of carbons for about thirty minutes

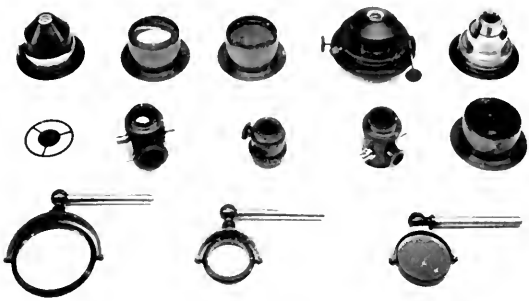


Fig. 7.—Condensers and illuminators used in microscopy

only. Frequently exposures are made at high powers (6,000 to 9,000 diameters) lasting from 45 seconds to 3 minutes, during which time the carbons may feed several times and no ill effects result, so perfect is the operation.

Critical illumination is nothing more than bringing the rays of light from the source of illumination to a state of proper focus and optical alignment so that the surface of the specimen under examination will be uniformly and brilliantly illuminated. This matter of securing uniform illumination is by no means the simple operation that the designation implies since usually an optical train consisting of the light source, condenser, diaphragms and an object illuminating device of some sort must all be brought into exact optical alignment with the optical system of the microscope.

For very low-power work or for gross photography of specimens, a gas-filled, metal filament lamp with a suitable condenser and mounted on a portable pedestal which may be adjusted to all angles is very useful. In this case the optical train is dispensed with and the light thrown at the proper angle on the specimen to be photographed.

Great brilliancy is not required for this work but rather a diffused light, obtained by means of interposing a ground glass screen in the illuminating beam.

The object in photomicrography is to record as clearly and as faithfully as possible the structural characteristics of the specimen. This is accomplished by a rendering of contrast between the structural elements of the specimen and by intensifying or diminishing this contrast to suit the particular characteristics which are to be reproduced to best advantage. This control of contrast is obtained by control of the color of the light used for illumination.

A spectroscopic analysis of the light of the arc shows a continuous spectrum consisting of three dominant color portions, blue-violet, green and red which pass by gradation to each other; the blue-violet passes by blue and blue-green to green, and the green by yellow and orange to red.

If an object absorbs some constituent of the white light falling on it then the reflected light will be deficient in this color and as a result the eye will experience the sensation of color.

The effect on the color of the residual light by blocking out a narrow band at different positions in the spectrum is shown in Fig. 7a.

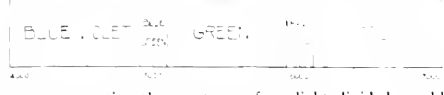


Fig. 8—Diagram representing the spectrum of arc light divided roughly into three dominant bands.

A simple diagrammatic representation of the visible spectrum is shown in Fig. 8, in which the tri-color division is broadly made as follows:

Blue-Violet	4,000 to 5,000 A.U.
Green	5,000 to 6,000 "
Red	6,000 to 7,000 "

An object which appears red to the eye when illuminated by white light is absorbing the blue-violet and the green light, and the bulk of what it reflects or transmits is red. Similarly, an object appears green because it is reflecting or transmitting the green constituents of the spectrum and absorbing the red and the blue-violet rays. These are simple cases assuming sharp absorptions and ideal conditions, but in the practice of the art of photomicrography we are dealing with gradation in color and oftentimes the structural characteristics

of the specimen show little contrast, either within the specimen itself, or between the specimen and the background. Therefore, to reproduce an object faithfully or to accentuate faintly revealed characteristics, careful consideration must be given to the color of the light used when photographing the specimen. For the purpose of separating white light into well defined bands, light filters are used and their function is to filter out rays or bands of rays of certain given wave-lengths. These filters consist of colored gelatine films mounted between flat pieces of glass or of liquids appropriately colored and contained in rectangular vessels of glass with flat and

Fig. 9 — Hilger wave-length spectrometer. The camera is interchangeably mounted with a reading telescope.

parallel side walls. The "Wratten M" series of gelatine and glass filters is probably the best known and most widely used. The selection of a light filter for a given specimen is usually by experimental methods. Successive filters are inserted in the illuminating beam and the resulting image studied for rendering and definition. However, two simple rules apply generally; if a color is to be rendered as black as possible, then it must be photographed by light of wave-lengths within the absorption band of the specimen; when contrast is desired within the specimen itself, the object should be photographed by light of a wave-length which it transmits. The first rule is of use when it is desired to secure contrast between the object and the background; and the second for better rendering of detail within the object.

Spectrometers are available for determining the characteristics of filters; for determining the transmission spectrum of a micro-

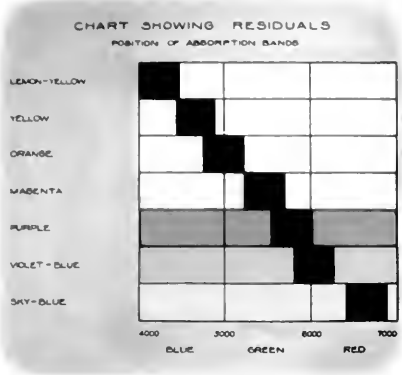


Fig. 7a—The effect on the residual color of arc light by blocking out narrow bands at different positions in the spectrum.

Fig. 10—Direct color photography of the spectrum of arc light and with the Wratten A, B, C and D filters respectively interposed in the beam.

scopic mount; and for studying the effect of dyes or stains on certain types of transparent mounts. For the purpose of filter studies the Hilger wave-length spectrometer constitutes a very useful accessory. This instrument is illustrated in Fig. 9, and in Fig. 10 is shown by direct color photography the residual light from an arc lamp after passing through various filters. The spectrometer is adapted for either direct vision work or photography, a camera and telescope being interchangeable. Instruments for observation with spectroscopically decomposed light constitute what are known as spectroscopic eyepieces and are very useful for certain classes of work, since they replace the usual microscopic eyepiece and may be used with any objective. Precision instruments of this type are capable of measuring the transmission or absorption spectrum of very minute

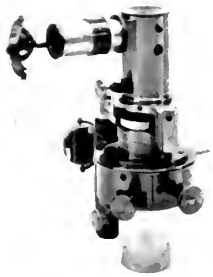


Fig. 11. A spectroscopic eyepiece by Zeiss. This instrument replaces the usual eyepiece of the microscope when it is desired to make observations with spectroscopically decomposed light. It yields an image of the transmission spectrum of the object with a superimposed Angstrom scale and if desired the transmission spectrum of the staining reagent may also be brought into the field of vision. The staining reagents are placed in glass vials.

bodies such for example, as a single blood corpuscle, which state of perfection is said to be attained by the Zeiss instrument, illustrated in Fig. 11.

The wave-length of the light used in photomicrography also has other useful functions to perform and for some classes of work these take precedence. Mention has been made of the correction of objectives for aberrations which are inherent in the simple lens. When an objective, not fully corrected, is used for photomicrography at the higher magnifications, color distortions assert themselves and result in faulty performance of the objective unless filters are used

to exclude light of wave-lengths other than that for which the objective has been computed.

In high-power photomicrography of metallurgical specimens, the purpose, of course, is to attain the maximum of resolution and here the wave-length of the light used plays an important part. As mentioned above the resolving power of an objective may be increased by decreasing the wave-length of the light used. Assuming that a Wratten "F" filter is used whose transmission band is from 6,100 A.U. to the red end of the spectrum, then an objective of 1.4 N.A. should resolve about 109,000 lines per inch. If a "C" filter is used whose spectral transmission is from 4,000 A.U. to 5,100 A.U., the same objective should resolve about 158,000 lines per inch. In other words, by using the shorter wave-length light,[1] it is possible to effect a theoretical improvement of about 45% in the resolution. In practice, these theoretical values are not fully obtained because of other complications entering into the problem.

Polarized Light

Polarized light is oftentimes a very useful aid in the study of transparent objects. By combination with suitable selenite plates color combinations are developed in the specimen and between the specimen and the background which facilitate identification of substances, comparison of known and unknown substances, and the study of their structure. In the field of crystal studies, polarized light is indispensable and it furnishes evidence of a very substantial nature in the field of micro-chemistry. The problem has been presented on occasions to identify the nature of some substance, resulting from the corrosion of some small telephone part. The evidence in these cases could easily be placed on the head of a pin but by the use of polarized light in conjunction with micro-chemical methods, it has been possible to form some sort of a qualitative estimate of the nature of the substance. Polarized light is obtained by means of a nicol prism contained in a suitable mount which is clamped in a ring beneath the sub-stage condenser. The illuminating beam from the microscope mirror is thus polarized before it reaches the condenser. A second nicol prism called the analyser is either contained within a special eyepiece or the analyser takes the form of a mount which may be placed above the usual eyepiece. Both polarizer and analyser are

[1] Regarding the use of ultra-violet light see High-Power Photomicrography of Metallurgical Specimens, F. F. Lucas, Trans. Am. Soc. for Steel Treating; Vol. IV, p. 611, 1923.

mounted so that they may be revolved. Extinction angles are read from a suitably graduated circle usually forming a part of the analyzing eyepiece.

Preparation of Specimens

Specimens, to be investigated or studied by microscopic methods must have a preparatory treatment in all cases except, perhaps, for very low-power work. Many samples require the preparation of transparent sections: that is, a specimen of the object a few thousandths of a millimeter in thickness so that it is transparent or at least translucent; studies of woods, porcelains, papers, fibers, tissues, insulating compounds, etc., are usually made with transparent sections.

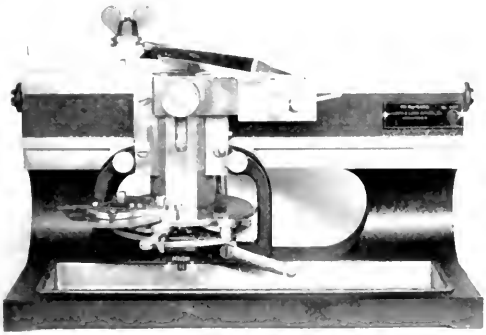


Fig. 12—A sliding microtome for cutting microscopic sections. The work is held in a clamp and a very heavy section razor, flat on one side and hollow ground on the other is operated backward and forward on a slide rails. The return movement of the razor operates the elevating mechanism to which the work is attached so that the latter may be raised to cutting position by predetermined increments.

Hard specimens such as porcelain are prepared by grinding, softer materials such as wood sections are first prepared by suitable softening processes and then are cut in an instrument called a microtome, a form of which is shown in Fig. 12.

Delicate structures require strengthening before they can be cut; these are embedded in paraffin, celloidin or glycerine gum. For successful results gradual and thorough impregnation of the parts is required and this operation may take several weeks. After the

sections are cut, they must be further prepared by being stained, dehydrated and cleared after which they are finally mounted in Canada Balsam or similar mounting medium between a glass microscope slide and a cover glass. Mounts of this kind are permanent, but when it is not desired to retain the mounted specimen for record or future examination, temporary mounts are often made in which the mounting medium is some liquid such as water or glycerine, or

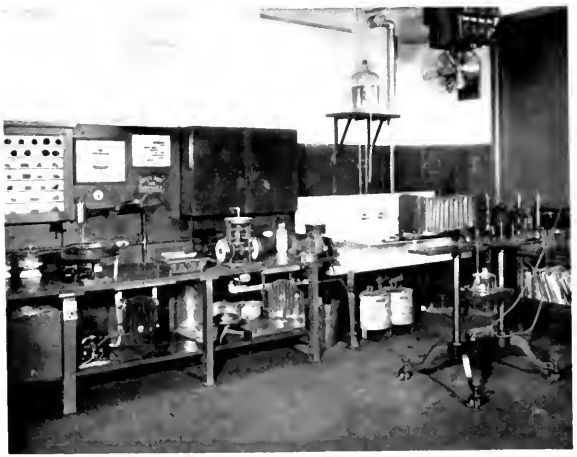
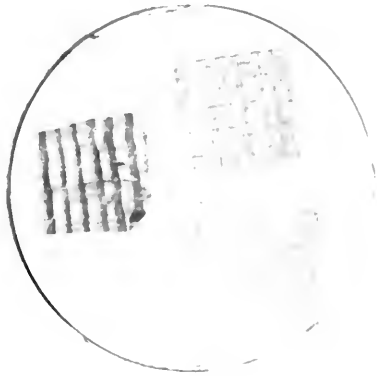


Fig. 14 Equipment for the preparation and preliminary examination of opaque specimens.

in some cases, may be the staining medium itself. An enlarged view of a permanently mounted transverse radial and tangential sections of Douglas Fir wood is illustrated in Fig. 13.

The preparation of metallurgical specimens is accomplished by different methods and if a specimen is to be examined at extremely high powers, the utmost in skill and refinement of methods is necessary. The usual method of procedure is first to file a flat surface on the specimen, after which the surface is gradually brought to a semi-polished condition by rubbing the specimen on a sheet of French emery paper, placed on a plane surface. A coarse grade of paper is first employed and by gradual steps, finer and finer grade papers are used, the rubbing on each successive paper being in a direction at right angles to the preceding paper and continued until the scratches



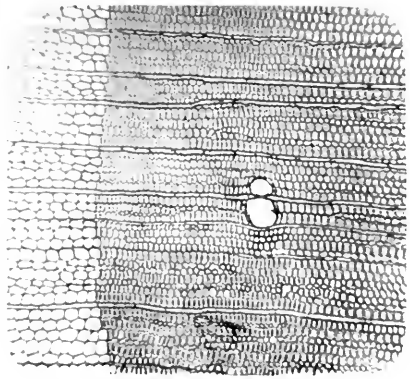


Fig. 13 An enlarged view of a specimen prepared for microscopic examination. The cover glass shown by the circle measures ¾ inch in diameter. The mounts are transverse, radial and tangential sections of a wood specimen and were stained to make their structure visible under the microscope. The appearance of a transverse section of Douglas Fir at 100 diameters is shown in the lower illustration.

of the preceding operation have all been removed and finer ones established in the new direction. This is continued to the 000 paper, after which the specimen is further polished on a polishing machine having a broadcloth covered lap capable of being revolved at varying speeds to about 1,200 rpm. This lap is kept moistened with water and fine alundum is used as the abrasive. This operation gives a

Fig. 15—General view of the Laboratory for Technical Microscopy.

semi-polish and when properly carried out, leaves the specimen with numerous very fine scratches. The final operation is carried out on another lap covered with very fine broadcloth and with an exceedingly fine abrasive such as the finest jeweler's rouge or magnesium oxide. For high-power work magnesium oxide is the only polishing medium which has been found to yield a satisfactory surface. The technique for the development of surfaces at high powers has been worked out in our laboratory so that it is now possible to study metal structures with great clearness at high powers. Equipment for grinding and polishing specimens is shown in Fig. 14.

Metals, after polishing, as a rule, do not show their structural characteristics, but must be treated in some way to etch the polished surface. This etching operation is a simple matter for low-power work, but as the magnification is carried higher and higher, the problem becomes increasingly difficult.

Bell System Photomicrographic Laboratory

A general view of the Laboratory situated on the fifth floor of the building at 463 West Street, New York City, is shown in Fig. 15. Some of the equipment is more fully illustrated in detail views.

It consists of two metallurgical equipments, one of which is the large Zeiss metallographic outfit shown in the foreground of Fig. 15. This equipment is of precision quality and is used for all classes of

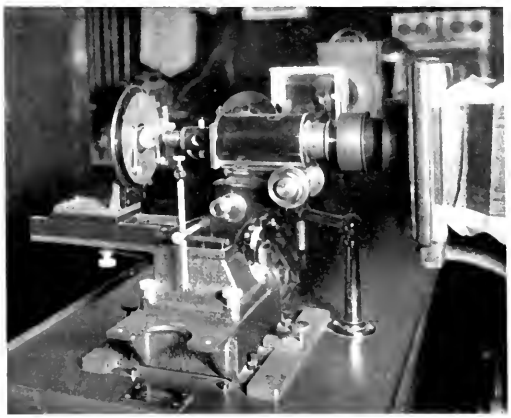


Fig. 16—The Martens stand of the large metallographic outfit. The vertical illuminator is shown between the barrel of the microscope and the objective.

work involving opaque specimens. The optical parts consist of a full complement of Zeiss apochromatic objectives and compensating eyepieces for medium and high-power work. For low-power work a full set of Zeiss micro-planar lenses and a Tessar lens are used. The maximum bellows extension of the camera is 155 centimeters and the plate holders are designed for 21 x 30 centimeter plates and all smaller sizes by employing suitable kits. The illuminating train consists of an automatic arc lamp, a condensing system, and cooling cells, mounted on an optical bench and capable of adjustment to meet the conditions of the work.

Illuminators of conventional types, for vertical and oblique light may be assembled on the Martens type stand. This stand is a departure from the construction of the usual form of microscope stand. It is much more rugged and is arranged for use in a horizontal

position only. In precision work a stand must be stable and substantial and the construction throughout has been arranged with this thought in mind. The microscope is equipped with a movable stage for rough focusing and this is fitted with a revolving mechanical stage so that the specimen while under examination may be moved about at will for the purpose of study or exploration. To facilitate focusing,

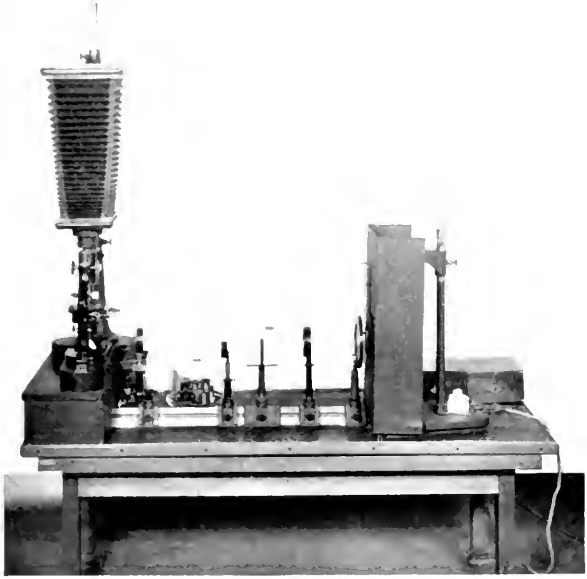


Fig. 17.—A vertical photomicrographic camera for transparent specimens.

gear is provided so that the operator may sit at the ground glass screen and by means of a wooden handle, focus the microscope. A ground glass screen for viewing and for rough focusing and a clear glass screen for fine focusing with a magnifier are provided to be interchangeably mounted with plate holders on the camera back.

A second metallurgical outfit of Bausch and Lomb manufacture shown in Fig. 14, is used for preliminary examination of specimens while in the course of preparation and for photographing some metallurgical specimens at medium powers. This outfit is also arranged

for photomicrography and has a 5 x 7 camera of rather short bellows extension. The objective equipment is of the achromatic type.

For transparent work a Zeiss vertical camera outfit, Fig. 17, equipped with the conventional Zeiss research type microscope is used. The camera has a bellows extension of 80 centimeters and uses 5 x 7 or smaller plates. It is fitted with ground and clear glass focusing screens similar to the large Zeiss metallurgical outfit. The illuminant is a 500 watt metal filament nitrogen filled bulb with the filament mounted so that a large circular area of illumination is presented, or if desired, the filament assembly may be turned sideways and a single

Fig. 18—A 500 watt metal filament, gas-filled lamp for use in photomicrography.

filament strand is thus presented to the optical train. In medium and high powers, this approximates a point source and for the lower powers the large circular arrangement of the filament provides a relatively large area of illumination which is quite desirable. The lamp is illustrated in Fig. 18. The illuminating train consists of condensing and cooling units adjustably mounted on a substantial optical bench as in the case of the metallographic outfit. The objectives consist of a full set of apochromats and also several achromats of low power. The micro-planars are also used with this equipment.

The Ultra-Microscope

The ultra-microscope is an instrument for revealing the presence of very minute bodies present as colloids in transparent solids or liquids. The presence of these particles is made apparent by the light rays which they intercept and diffract upward into the microscope objective. It is a matter of common observation that dust particles are seen in an intense beam of light such as sunlight but otherwise their presence remains concealed. This principle of illumination is made use of in the ultra-microscope as described below and accordingly differs considerably from the conventional arrangement of compound microscope and illuminant.

The appearance of ultra-microscopic particles in fluids and transparent solids as seen by means of the ultra-microscope is, without a

doubt, one of the most fascinating and spectacular demonstrations within the scope of technical microscopy. A beaker containing water with a drop or two of glue or soap, or containing benzol with a few drops of a rubber solution stirred into it, or even some rather dirty looking oil which has seen service in some machine, do not constitute

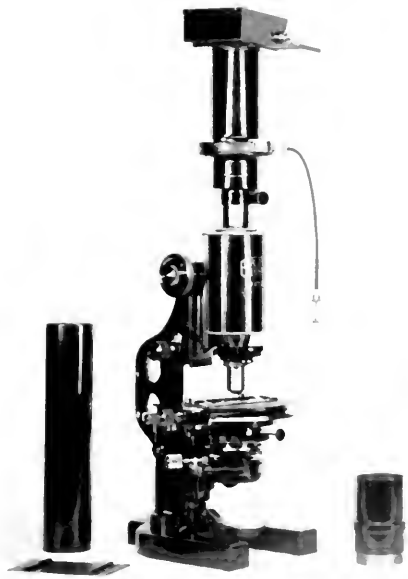


Fig. 19.—A small photomicrographic camera developed by the writer and used extensively in the laboratory for photographing on film, or on plates. It is used when a large number of small specimens are to be reproduced or when a large field is unnecessary.

interesting exhibits as viewed in the beakers, but placed in suitable cells for ultra-microscopic examination, these liquids come to life and display the colloidal particles coming into vision as tiny illuminated particles, only to burst into rings of light and pass away into the dark background. The constant irregular motion is the Brownian movement and the smaller the particle the more lively it moves. Conglom-

erate masses of particles merely float through the field of vision and, compared with the individual particles, appear exceedingly sluggish.

Fig. 20 gives a general view of the Zeiss ultra-microscope as originally devised by Siedentopf and Zsigmondy. The equipment has been superseded to some extent by the later Siedentopf cardioid ultra-microscope. The latter is a very powerful light-concentrating device

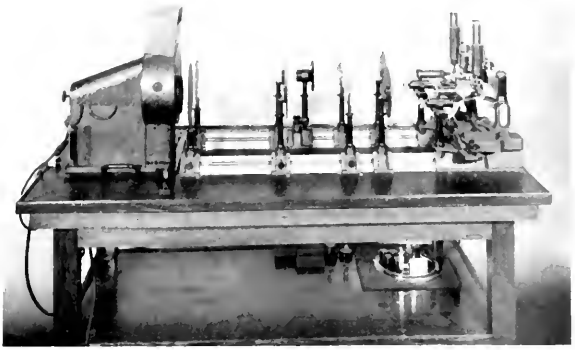


Fig. 20—The Slit ultra-microscope for transparent solid or liquid specimens.

and for this reason it is primarily adapted for the examination of fine colloidal solutions and dilute precipitates as well as for the observation of micro-chemical and photo reactions. For transparent solids and for the precursory examination of liquids and for rapidly passing in review several fluids in succession, the original arrangement retains marked advantages. The cardioid ultra-microscope will be described more fully later on.

Fig. 21 shows diagrammatically the path of the rays within the preparation in the presence of ultra-microscopic particles and will serve to make clearer what is to follow. In the original form of ultra-microscope (Fig. 20) the horizontal incident rays which go to furnish the illumination do not enter the microscope, the latter being set up vertically and hence the background appears dark. The only rays to enter the objective of the viewing microscope are the diffracted rays which come within the aperture of the objective.

At one end of the base board is an automatic arc lamp mounted on slide rails so that it may be brought in line with either of two illuminating trains mounted on optical benches.

One of these illuminating trains functions with a microscope which has mounted on its objective a clamping device for holding Biltz cells in which the liquids are placed for examination. The other train serves another microscope on the stage of which is mounted a special object stage capable of being raised and lowered and provided

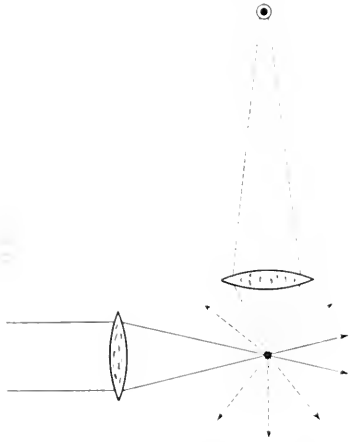


Fig. 21.—Illustrating diffraction of light impinging upon an ultra-microscopic particle. Illuminating rays represented by solid lines and diffracted rays by dotted lines.

with a plate at the top to receive the specimen to be examined. In this case the specimen, if a hard solid, has been previously prepared to have two ground and polished surfaces in planes at right angles to each other and is mounted so that one faces the illuminating train and the other the objective of the viewing microscope. Plastic substances or certain liquids not suited to the use of the Biltz cells are placed in a special glass cell having a deep cylindrical recess faced with a quartz window toward the illuminating train. Various cells for ultra-microscopic examinations are shown in Fig. 22.

Placed next to the arc lamp is a fixed diaphragm and then a small projection lens which is corrected chromatically and spherically and brings the image of the positive carbon of the arc lamp to a focus on the adjustable slit. The slit is provided with a drum bearing a scale. The divisions of the scale embrace 50 parts and a complete revolution of the drum opens the slit $\frac{1}{2}$ mm. so that each division of the scale advances the slit $\frac{1}{100}$ mm. The slit is fitted with two jaws at right

angles to the principal slit, one being adjustable by a milled screw head. The function of these jaws is to limit the length of the slit. The slit head may be given a quarter turn so that it may be set horizontally or vertically, which is necessary in order to calibrate the instrument as explained later. A projection lens next in order toward the microscope projects the image of the slit into the image plane of a horizontally mounted objective which is mounted on a stand with cross

Fig. 22—Cells used for the examination of fluids with the ultra-microscope.

slides so that the objective which serves as an illuminator may be moved horizontally in two directions, at right angles to each other. The movement of the cross slides is controlled by screw adjustment but for coarse adjustment in the direction of the illuminating train a sliding sleeve adjustment is made. By this means the illuminating objective can be centered with respect to the observing microscope objective. In the correct position the front lens of the illuminating objective is about 1 mm. from the mount of the observing objective.

The Biltz cell has a rectangular cross section which permits of accurately adjusting the cell in position. A thistle funnel at one end is for the reception of the liquids; the other end is provided with a piece of rubber tubing which has a pinchcock to prevent the escape of the fluid. The rectangular section of the cell has two quartz windows, one of which normally faces the illuminating objective and the other the observing objective. The cell is attached to the observing objective by means of the clamp mentioned and the cell is focused in the proper position in the beam of light by racking the microscope draw-tube upward or downward in the usual manner by the coarse and fine adjustments. The observing objective is a special water immersion objective which makes contact with the upper window of the Biltz cell through the medium of a drop of distilled water.

Quantitative investigations are made by counting the visible particles in a given volume of the fluid and the manner in which so novel an investigation can be accomplished by optical means should prove of general interest. One method consists in the use of the eyepiece micrometer which is substituted for the ordinary eyepiece of the observing microscope. The eyepiece micrometer is ruled into squares and the dimensions of these are found by calibration with a stage micrometer. The depth of the stratum is measured by turning the slit head through a right angle and thus a solid is blocked out in the path of the light rays, whose length and breadth are defined by the rectangular area of the micrometer eyepiece and whose depth is

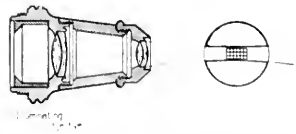


Fig. 23.—Illustrating the adaptation of micrometry to the ultra-microscope for the purpose of counting particles per unit volume.

that of the light beam and may be read from the known dimensions of the eyepiece micrometer. Fig. 23 shows the cross ruling of the eyepiece and the pencil of light which traverses the field. The length of the side of each square as seen through the water immersion objective with a tube length of 160 mm. has an approximate value of 9 μ as referred to the object, which value is sufficiently accurate for ordinary measurements. Where more exact measurements are required, the ruling is calibrated for the eyepiece and objective by means of a stage micrometer in the manner to be described under the subject of micrometry.

For studying the behavior of particles in polarized light the eyepiece is fitted with an analyser. In a measure, as the particles decrease in size they become more strongly polarized in degree towards the plane which passes through the axis of the illuminating and diffracted rays, i.e., the principal plane of diffraction. The analyser also serves to distinguish unpolarized from polarized light.

The apparatus for examination of solids is identical in so far as the illuminating train is concerned with the apparatus for liquids just described. It differs only in the character of the specimen or of the cell used and, while designed primarily for transparent solids, it may be used with a suitable cell for liquids also. When liquids are being examined, there is no need for searching the specimen since

the particles are in constant motion, but when solids or semi-solids are being examined it may be desirable to do so. The mechanical stage of the microscope on which is mounted the adjustable specimen stage allows any layer in the specimen to be brought into accurate focus and hence various strata of the specimen can be examined one

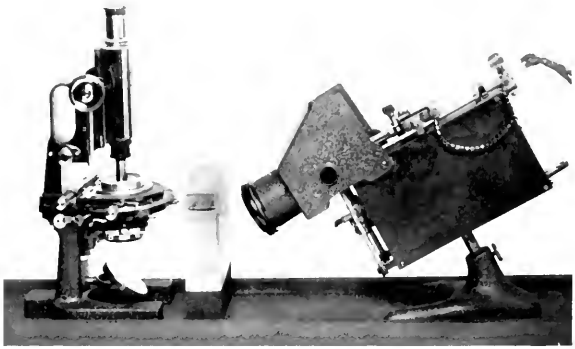


Fig. 24—Cardioid ultra-microscope.

after the other. As previously stated, the specimen must be provided with two polished surfaces at right angles to each other to correspond to the quartz windows of the Biltz cell.

Since the observation of ultra-microscopic particles in polarized light supplies useful information respecting their form and color, a polarizer is provided with a hinged stand so that it may be swung out of the optical train. The analyser, as previously mentioned, is fitted over the eyepiece of the microscope.

The cardioid ultra-microscope illustrated in Fig. 24 differs only in two important features from the ordinary form of microscope. The illumination of the fluid under examination is obtained by a dark-ground condenser mounted in the sub-stage condenser collar and to which Zeiss has given the distinctive name "cardioid condenser." A diagram of the condenser and the paths taken by the rays is illustrated in Fig. 25. Since the aperture of the rays brought to a focus by the condenser exceeds 1.0, it follows that no light can emerge from the condenser if there be a stratum of air above the condenser. It is therefore necessary to connect the object slide

or cell chamber and the top of the condenser by a stratum of immersion fluid free from air bubbles. Cedar oil or pure water is used for this purpose. The chamber for the cardioid condenser is illustrated in Fig. 22. The chamber is made of quartz glass and consists of a circular disc having on one side a circular groove and an optically plain central portion within the groove about 2 μ below the plane outside the groove. A drop of the fluid to be examined is placed on this depressed central portion and a cover glass of quartz placed over it. The excess fluid is expelled to the annular groove and a stratum about 2 μ in thickness is retained in the central portion of the chamber for microscopic examination. The cell is assembled in the metal mount which has a clamping ring and a recessed member

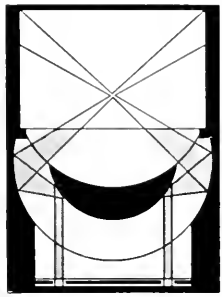


Fig. 25—Diagram of the rays in a cardioid condenser.

to receive it. The very brilliant illumination resulting from the cardioid condenser would cause glass to fluoresce and for this reason a quartz cell is used. Moreover, glass is more liable to be affected by corroding agents than is quartz.

The utmost care must be taken to prepare the cell chamber. This includes washing with alcohol and water; dipping in boiling sulphuric and chromic acid solution; washing in tap water; rinsing in distilled water and then in redistilled alcohol; drying in a hot air current and finally cooling under a bell jar; all of which is necessary to insure absolute cleanliness.

An automatic arc lamp is used as a source of illumination and the image of the crater is projected by a projection lens onto the mirror of the microscope from which the rays are reflected upward into the cardioid condenser.

The objective used with the cardioid condenser is a special apochromat 3 mm. 0.85 N.A. glycerine immersion lens which constitutes a homogeneous immersion lens for cover glasses of fused quartz. This type of objective is necessary because the success of the observation is then largely independent of impurities and slight blemishes on the upper surface of the cover glass, moreover, the lens confers a greater immunity from the effects of varying cover-glass thickness and the immersion fluid precludes the entrance of dust which would gradually cloud the image.

Slit ultra-microscopes are not arranged for photography because in the case of liquids the particles are in a rapid state of motion and the illumination is insufficient. Since in transparent solids the particles are stationary, the image seen in the slit ultra-microscope may be reproduced by making a lengthy exposure. With a small photomicrographic camera developed by the writer the image seen in the slit ultra-microscope for solids has been reproduced and, by instantaneous photography, the moving particles in liquids as seen in the cardioid instrument. Except for purposes of evidence or record, there is little to be gained by photographing with the ultra-microscope.

Dark-Ground Illumination

The dark-ground illuminator constitutes another aid to microscopic investigation. This, in reality, is a sort of ultra-microscope, since the objects are viewed by diffracted light much in the same way as in the cardioid type of equipment. This method of illumination is accomplished by stopping out the axial rays and allowing those of greater aperture to strike the specimen at an angle. The usual form of condenser may be made to yield dark-ground illumination by the simple expedient of inserting a central stop in the path of the light rays just below the sub-stage condenser in a ring provided for such purpose. Better results are attained by use of dark-ground illuminators which are special condensers designed with this object in mind. Dark-ground illumination furnishes valuable means for bringing into view objects which are smaller than about 1 μ. Examples of such objects are furnished by fibers, fine crystalline needles, fissures, edges, rods, bacteria, etc. Under dark-ground illumination methods, these objects are easily seen and studied, whereas with transmitted light, they can be seen with difficulty unless rendered distinguishable by staining. Certain bodies with laminar markings are also suitable subjects for dark-ground studies and in this case the markings are distinguishable more by reason of dissimilarities in refraction than by differences in coloring.

Micrometry

Micrometry plays an important part in technical microscopy because the dimensions of micro-constituents in a specimen are helpful for purposes of identification or for forecasting physical properties. In metallography the measurement of grain size is assuming importance for certain alloys and in some cases, specifications are so drafted as to define this characteristic.

For measuring objects under the microscope, the eyepiece contains a glass disc on which fine divisions have been ruled. In some cases, these rulings take the form of a cross-section composed of small squares or rectangles. The reading of each division of the eyepiece micrometer is calibrated for each objective by comparison with a standardized stage micrometer. These stage micrometers are glass microscope slides on which known units of length have been accurately ruled, such as 1 mm. divided into one hundred parts or 3 mm. divided into tenths and one-tenth divided in hundredths of a mm., etc. The stage micrometer is focused in the same way as a microscopic specimen and adjusted into position so that the rulings of the eyepiece micrometer are superimposed on them. It is then possible to evaluate the eyepiece rulings in terms of the standardized stage micrometer, after which the latter is removed and the specimen to be examined substituted in its place. Thereafter, the image of the eyepiece rulings will be superimposed on the image of the specimen and measurement can proceed. For precision work, a very accurately made eyepiece micrometer is used, a typical form of which has a thin glass plate upon which is ruled a cross and a double line. This is mounted on a slide immediately below a stationary micrometer scale and can be moved by means of micrometer screw. The cross is accurately set by the micrometer screw to coincide with the particle to be measured, the double line serves to count complete revolutions of the screw with the aid of the scale which is seen in the field of vision. The screw carries a drum which has 50 divisions and each division corresponds to a displacement of the cross through a distance of 0.01 mm. so that a complete revolution of the drum displaces the cross 0.50 mm. The actual readings of each interval of the drum head must be accurately calibrated for each objective by means of a stage micrometer. A group of instruments for use in micrometry is illustrated in Fig. 26.

Applications of Photomicrography

In closing, attention is directed to the photomicrographs comprising the Appendix of this paper, each of which was taken in connection with

some definite engineering problem involving telephone apparatus. As the useful range of microscopic vision is extended farther and farther into the realm of higher magnifications, a more exact knowledge of materials is obtained and the effect is learned of physical and chemical forces acting to destroy or to build.

It has been conceded quite generally that about 1,500 diameters of magnification represents the limit of useful magnification. As previously stated this is a much disputed question. Laboratory studies, painstakingly carried out over a period of several years, have

Fig. 26—Various types of eyepiece and stage micrometers used in connection with the microscope to obtain the dimensions of microscopic objects.

accomplished improvements in technique and in precision of adjustment of the equipment which have shown that remarkable resolution, depth of penetration and clearness can be attained in the case of metallurgical specimens, at extremely high powers. There seems little reason to doubt that our knowledge of metals can be augmented very materially by studies of their structures at high powers.

Moreover, it seems probable that the finest high-power objectives are of a quality beyond our ability to use them to best advantage because of our incomplete knowledge of how best to prepare specimens for examination at high powers.

It is impressive to evaluate magnification in terms more readily comprehended. For instance, the cross section of the average metallurgical specimen may be considered as a square whose side measures one-half inch. If we magnify this specimen 100 times, obviously we have an area measuring 50 inches on the side, but if we magnify it 10,000 times, then we have the equivalent of an area about 415 feet on a side or roughly, about four acres. An average picture at 6,000 diameters is 6 inches in diameter and therefore by a reverse order of reasoning, the actual area of the specimen under observation becomes 1 1000 inch in diameter.

APPENDIX

Fig. A. Meteoric iron consists of iron, nickel and the other elements usually found in steels such as carbon, sulphur, phosphorus, etc. The study of meteorites has contributed much valuable knowledge to the science of metallography. The Widmanstatten figures (shown by the arrangement of the constituents with reference to crystallographic planes) were generally considered characteristic of meteoric iron and it was believed that they were not to be found in manufactured iron and steel. Later this was shown to be an incorrect view.

(a) A meteorite which fell at Carthage, Tenn., containing 89.46% iron and 7.72% nickel and which shows the octahedral Widmanstatten structure. Magnification 4 X.

Fig. A

(b) Meteoric Crystals. The figures are sections through an octahedron and were developed by suitably etching a polished surface of the meteorite. Their perfect form is indicative of very favorable conditions of growth and is a corroboration of the octahedral crystalline form of the meteorite. Magnification 3500 X.

Fig. A

(c) A cast steel of 0.5% carbon in which the Widmanstatten or cleavage structure has developed somewhat similarly to that shown in the meteorite. The physical characteristics of the steel are dependent on the structural arrangement of its constituents, in this case pearlite (dark) and free ferrite (light). By suitable heat treatment this coarse structure may be refined and the physical properties of the steel greatly improved. Magnification 100 X.

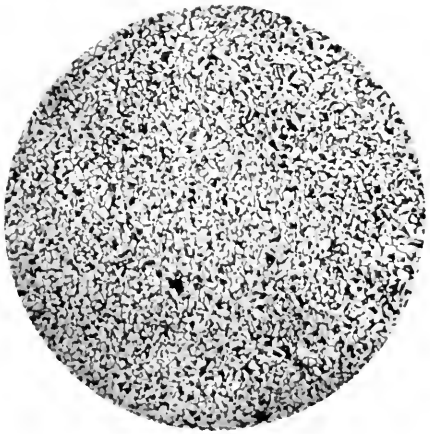


Fig. A

(d) The same steel as illustrated in "C" but after being refined (heated to 1000°C; air cooled; reheated to 650°C, and again air cooled.) Magnification 100 X.

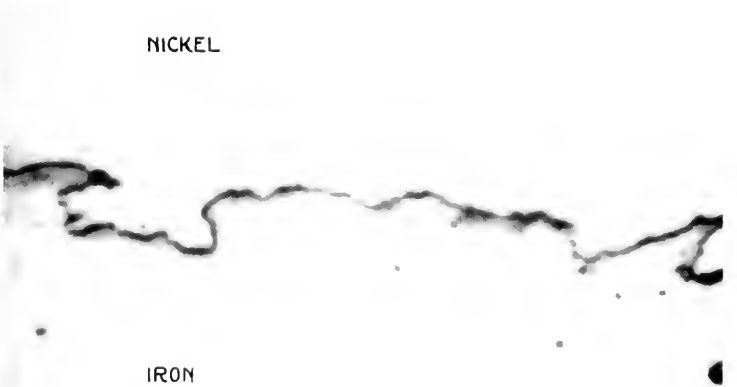
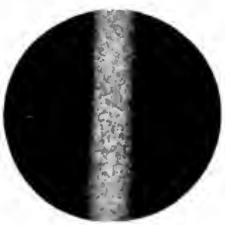


FIG. 13.—The application of high-power photography to the study of nickel finishes. One of the characteristics of an improved process for plating ductile nickel on iron is the interlocking or "keying" of the nickel and the iron. Magnification, 6000 X.

Fig. C. Distribution of filler particles in soft rubber insulation as revealed by a transparent section. The section was cut in a microtome by "flashing" the rubber with liquid air which hardened it just sufficiently to cut properly. The specimen was photographed by polarized light and with selenite plates to secure contrast between the particles and the embedding rubber compound. Note agglomeration of the particles into large masses. The ideal condition of distribution would be attained when each individual particle is surrounded by rubber. Magnification 720 X.

Fig. D. Colloidal particles as seen through the ultra-microscope.

(a) Polymerized particles in a phenolic resin solution. Taken with the cardioid ultra-microscope and the Lucas Photomicrographic camera. Instantaneous exposure was necessary because the particles were in constant motion. Magnification, 220 X.



(b) Coloring matter in glass. The glass was colored saffranin and was transparent to the eye or with any other method of microscopic vision but with the slit ultra-microscope the colloidal coloring matter becomes visible. Also taken with the Lucas photomicrographic camera, a time exposure being necessary. Magnification, 100 X.

Fig. E. Paper fibers by 97 X. Note the surface markings; the gradation in color and the appearance of roundness possessed by some of the fibers. The photograph was taken with a modern medium-power apochromatic objective.

Microscopic examination of textile and paper fibers affords a means of identification second to none. The fibers are recognized by characteristics peculiar to each and by color reactions to different stains. Cotton, for example, appears as a flat ribbon-like fiber twisted spirally; linen is round and shows "joints" and cross markings. The specimen illustrated consisted mostly of linen with a small proportion of cotton added.

Fig. F. Electrical porcelain by polarized light, magnification 100 X.

The quality of the porcelain may be judged to a considerable degree by a microscopic examination. The degree of vitrification is indicated by the rounding of the sharp corners on the quartz grains; whether or not the porcelain is homogeneous may be determined by the uniformity in distribution of the undissolved particles, and fissures, cracks, or voids are readily seen. All of these factors influence the physical characteristics of the porcelain.

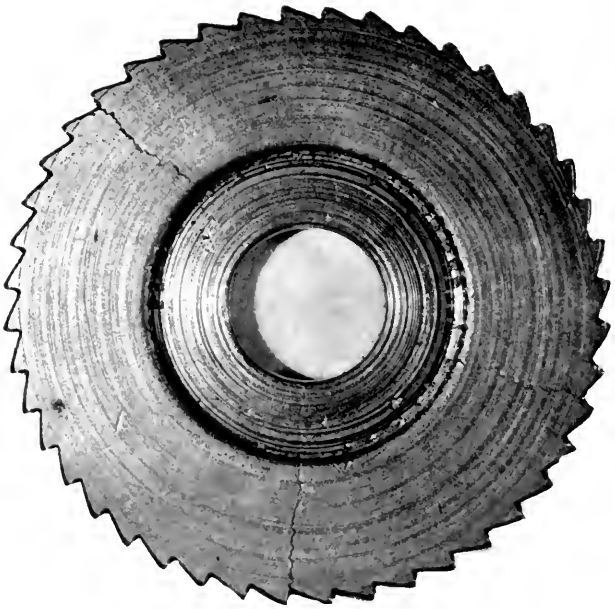


Fig. G. Season cracking of aluminum bronze ratchet wheels. (a) Ratchet wheel at magnification 2¾ X.

Fig. G

(b) Showing a large crack, at magnification of 23 X.

Fig. 6

(c) Intercrystalline nature of the cracks and the severely worked condition of the metal as indicated by several groups of slip bands traversing each crystal grain.

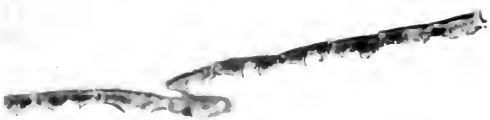
These ratchet wheels developed radial cracks while in storage or in service. Some of the cracks were so large as to be plainly visible to the unaided eye and others were of microscopic dimensions. They resulted from the metal being severely cold worked at the time the parts were machined and then left in a strained condition. The intercrystalline nature of the cracking is shown in "c" which is characteristic of season cracking. This illustration also shows the crystal grains traversed by several groups of slip bands, indicating the severity of the cold working.

Fig. H. Manila hemp rope is used extensively in telephone work and the fiber from old rope is used in paper for cable insulation which finds its way into the plant.

Microscopically the fiber is identified by certain characteristics, prominent among which are the silicified tabular cells known as stegmata. If the fiber is burned and treated with dilute acid the stegmata remain behind, resembling strings of beads.

Manila hemp makes the best cordage but it is somewhat difficult to distinguish the fiber from that of sisal which produces inferior cordage. The presence of the silicious skeletons of the stegmata and the color of the ash (grayish-black in the case of Manila hemp and white in the case of sisal) aid in the identification of the fiber.

(a) Manila Hemp Fibers, magnification 50 X.



(b) Ash of Manila Hemp, magnification 150 X.

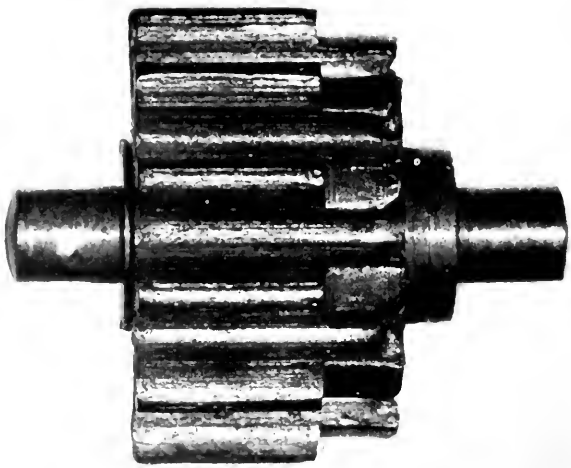
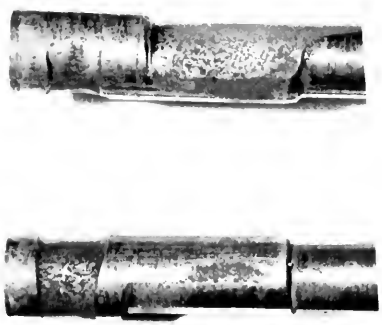


Fig. 1—A further illustration of low-power photomicrography in the study of telephone parts:
(a) Intermediate pinion of calling dial. Diameter of pivot .050 inch. Magnification, 14 X.



(b) The effect of laboratory wear tests on small shafts. Magnification, 4 X.

Fig. 1.—Case-Faulty pack hardening of cam wheels in roll ticket distributing system. The object of pack hardening is to impart a highly carburized wearing surface to the otherwise soft steel part. The interior remains soft and ductile. Lack of uniformity in hardening or insufficient depth of the carburized zone causes soft spots which result in unequal wear. The magnification is 11.2 X.

Fig. J

(b) Showing inappreciable depth of carburized zone, and a large non-metallic inclusion in the steel. Inclusions of this sort denote poor quality or dirty steel. Magnification, 100 X.

Fig. K. Steel of 1.5% carbon heated to 825° C. and quenched in oil. This medium power photomicrograph at 100 X really tells very little about the steel, except that it possesses a fine structure.

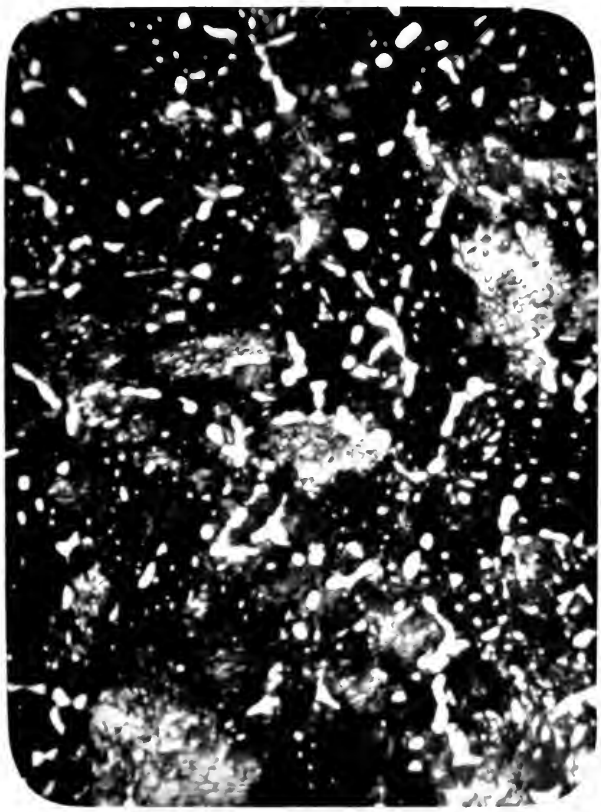


Fig. K (*continued*). At 2170 X the specimen is seen to be practically completely converted to spheroidized cementite. The cementite (iron carbide) which makes the light constituent is in the process of transforming from a laminated structure to small globules. Grain boundaries are still marked by accumulations of cementite but this is spheroidizing. In the light patches stratification [illegible] cementite is just visible.

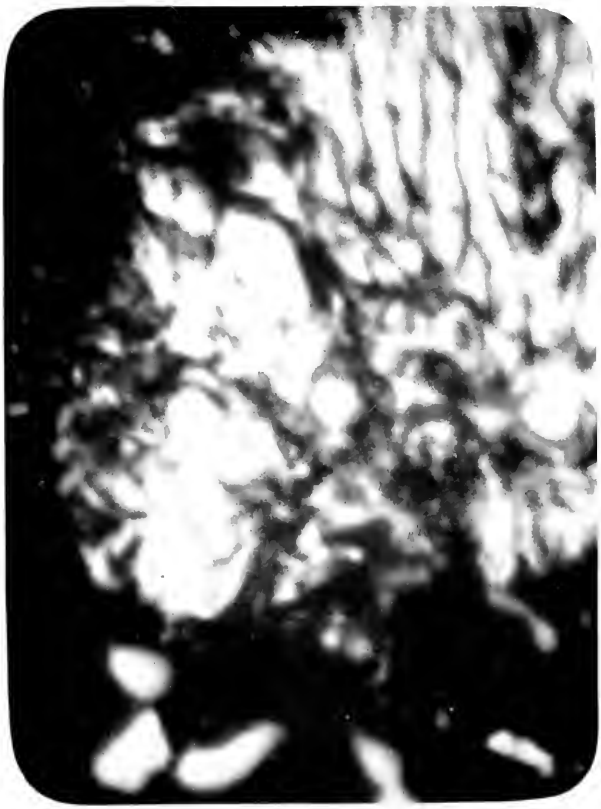


Fig. K. (*continued*) Under higher magnification one of these patches shows clearly the remaining vestige of laminated structure and the commencement of spheroidization. Magnification 9000 X.

Fig. L. Direct autochrome color reproduction of a stained specimen of Southern Yellow Pine showing sap-stain fungus mycelium. Magnification 100 X.

This particular fungus is harmful to the extent that it causes a discoloration of the sapwood, which assumes a blue color in place of the usual straw-yellow. Wood-destroying fungi differ somewhat in their appearance from the one illustrated.

Fig. M. Direct color photomicrography by the autochrome process of a radial section of mahogany wood. Magnification 50 X. Mahogany is one of the best of cabinet woods and finds wide application in the telephone plant.

A Clock-Controlled Tuning Fork as a Source of Constant Frequency

By J. G. FERGUSON

NOTE—The art of electrical communication employs such a wide variety of methods for the transmission of intelligence that it utilizes alternating currents whose frequencies cover the entire range between a few cycles per second and several million. With the increasing use of these methods, it becomes more and more imperative that determinations of the frequency of any alternating current may be made with extreme accuracy. In particular, recent developments in carrier current telephony and telegraphy over wires have placed exceedingly rigorous limits on the frequency adjustment of certain types of apparatus. It is many times necessary to hold such equipment as oscillators or filters to within 0.1 per cent. of given frequency values under commercial operating conditions. This means that calibrating devices used in the manufacture and maintenance of such circuits must be reliable to 0.01 per cent. and that the primary standard should be good to about 0.001 per cent. or one part in 100,000.

The present paper discusses one of the methods recently developed in the Bell System Laboratory for obtaining a source of practically constant frequency with which other frequencies may be compared. It consists of a clock-controlled tuning fork making 50 vibrations per second and, as is shown, the maximum deviation of its frequency from the mean is less than one part in 50,000.

A study has also been made of means for improving the constancy of the control clock and a new type of clock mechanism consisting of an electrically actuated pendulum, the impulse of which is controlled by a photo-electric cell, is suggested.—EDITOR.

INTRODUCTION

THE art of clock making is of such long standing that there have been few improvements of note in the last fifty years tending to increase accuracy. The average rate of oscillation of a good clock when taken over a sufficiently long period of time as, for instance, a day, can be held constant to about one part in 1,000,000. This accuracy is sufficiently high for all ordinary requirements in the measurement of time, including the field of electrical communication.

However, in electric measurements, the problems which present themselves ordinarily require the accurate measurement of intervals very much shorter than a second which is usually the smallest interval registered by the average clock. In solving these problems, we are therefore forced to the alternative either of designing a clock to have a period very much shorter than those of existing clocks or of using some form of short period oscillator whose uniformity can be controlled by the second impulses from a clock.

The first method has been admirably worked out as described by other members of the staff of this laboratory.[1] In this system a

[1] Paper by J. W. Horton, N. H. Ricker and W. A. Marrison, presented at the annual convention of the American Institute of Electric Engineers, June, 1923.

hundred cycle fork is kept in constant oscillation by a regenerative method, the conditions being so controlled that the mean period of the fork compares favorably with that of a good clock.

The attraction of the second method lies in the possibility of obtaining a sufficiently constant standard of frequency with nothing more than a good clock and standard auxiliary apparatus easily capable of application to any oscillating system. Such an outfit could be made available in cases in which the expense incident to the installation and maintenance of more elaborate equipment would not be justified.

Requirements of a Clock-Controlled Frequency Standard

It is a comparatively simple matter to control or operate a fork, or other oscillating system, by means of periodic impulses from a clock, so that the total number of oscillations will be some definite multiple of the number of impulses from the clock. However, the present requirements are more severe than this. It is necessary to have the oscillator operated so that each oscillation will be sensibly equal in magnitude and duration to every other oscillation. In other words, it is not sufficient that the clock and the oscillator keep in step over a given period of time, but the instantaneous frequency of the fork must not depart appreciably from the mean frequency. This requires a form of control which will not be to any extent discontinuous, but which will change uniformly in proportion to the divergence of the oscillator from the clock. Such a form of control in turn requires that the frequency of the oscillator itself be sufficiently constant when uncontrolled, to reduce all momentary fluctuations and rapid frequency changes to a minimum. This requirement is best satisfied by an oscillating system having a low decrement. Since a mechanical system is usually far superior to an electrical system in this respect, and since the most available mechanical oscillator for the range of frequency in question is a fork, our choice naturally falls on this form of oscillator.

A good fork maintained in continuous operation by some electrical means, such as regeneration, or a make and break contact and a driving magnet, is a comparatively simple system and is capable of a high degree of constancy.[2] It therefore satisfies all of the requirements for our purpose, but there remains the devising of some control which will be proportional to the divergence of the fork from the clock controlling it. In order to use any such control it is practically necessary to integrate the oscillations of the fork so that we may obtain a

[2] H. M. Dadourian, Phys. Rev. 13, page 337, 1919, "On the Characteristics of Electrically Operated Tuning Forks."

time interval equal to the number of cycles of the fork which we desire to make equal to the time interval between successive clock impulses. This is readily accomplished by means of a phonic wheel or synchronous motor operated by the fork. This motor may be connected to any form of gear train in order to get the necessary integration.

The requirements so far outlined do not limit the frequency of the fork in any degree except that we must be able to integrate its periods, and if a mechanical means is used as outlined, this probably sets an upper limit on the frequency at 400 or 500 cycles. However, practical considerations will generally make the most satisfactory frequency considerably lower than this, since it is an easier matter to compare unknown frequencies with a low frequency standard rather than with one of high frequency.

Method of the Control of the Fork by the Clock

The fork used in the system described below is of the same type as that tested by Dadourian. It is operated by a driving magnet and make and break contact, and was originally designed for use in multiplex printer telegraph circuits. It can be adjusted to operate at 50 cycles and is designed to drive a synchronous distributor which rotates once for every 10 cycles of the fork. By means of a 5 to 2 reduction gear and a contact operated by it, an impulse may be obtained once every 25 cycles of the fork. If the fork oscillates at exactly 50 cycles per second, the time interval between the impulses will be exactly one-half second, and this time interval will be shorter or longer, according as the speed of the fork increases or decreases.

The control system used is designed to affect the frequency of the fork in proportion to the difference between half second intervals as measured by the clock and the time required by the fork to complete 25 cycles. Fig. 1 shows the details of this control. Fig. 2 is the schematic diagram. Referring to Fig. 2 the contact marked "Fork" is the contact obtained every 25 cycles from the fork and the contact marked "Clock" is that obtained every half second from the clock. Each of these contacts is adjusted to remain closed for a period of approximately .05 second when operated.

The control operates as follows. When the clock contact closes, the relay operates and locks until the fork contact closes and short-circuits the winding of the relay which then releases. During the time that the relay is operated, the condenser C is charged through the resistance r_1 by the battery B_1. The voltage of this battery is such that when applied to the grid of the vacuum tube, it will just

reduce the space current to zero. The condenser C continuously discharges through the resistance r_2. The mean potential on the condenser is thus applied to the grid of the vacuum tube and modifies the space current, which, in turn, is passed through the damping coil

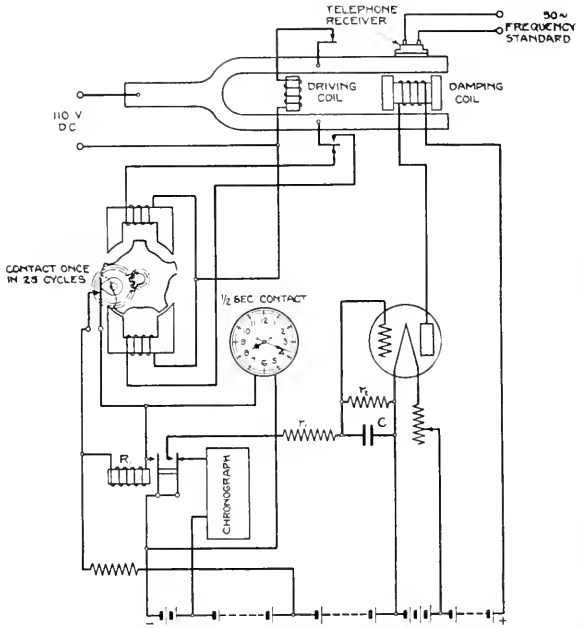


Fig. 1 Clock Control of Fork Frequency

of the fork. A stable condition is reached when the condenser discharge each second is exactly equal to the charge. Any variation in the condenser potential varies the current through the damping coil and changes the frequency of the fork. Now if the period, during which the relay remains operated, increases, the mean potential on the condenser will gradually increase. This will increase the mean negative grid potential, reducing the mean space current through the tube and through the damping winding, thus reducing the damping on the fork and increasing its frequency.

This method of control is slow yet sensitive to very slight changes in frequency.

The method of controlling the frequency of the fork by a damping winding was found to be the most simple and satisfactory method. The amount of variation of frequency which this winding will produce under extreme conditions should be slightly greater than the maximum variation to which the fork is subjected in operation when uncontrolled. This has been found to be about .05%, when temperature

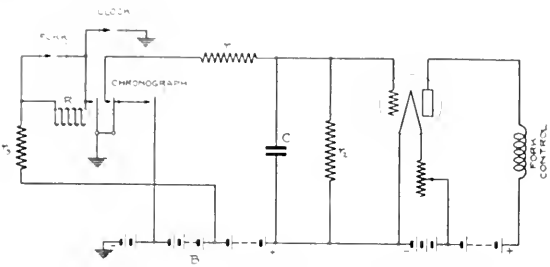


Fig. 2 (Clock Control of Fork Frequency. (Schematic Diagram)

variations are held to within a few degrees. The only requirement for the coil is that the current available from the vacuum tube must be sufficient to produce the necessary control. This requirement is easily met. It has been found possible with the equipment described to obtain an effect at least 10 times greater than necessary.

The change in frequency of the fork due to the current in the damping winding is a combination of several effects. The current will increase the decrement of the fork, due to the losses induced in the metal of the fork while vibrating in a magnetic field. This will cause a decrease in frequency. The magnetic force acting on the tines of the fork, even though it be assumed to cause no losses in the fork, is unsymmetrical, having a greater effect at the ends of the swing of the fork. This unsymmetrical force may also be shown to cause a decrease in frequency. Again there is a change in frequency due to the change in amplitude alone. For the type of fork here used this change may be an increase or decrease, depending on the range over which the change occurs.

It is obvious that when the control is operating, the voltage of the condenser and hence the space current of the vacuum tube, will fluctuate each half second. Since it is only the mean value of space

current that is used to control the fork, it is important that this fluctuation be reduced to a small amount. This may be done by using a large capacity C or a large resistance R_1. However, the effect of increasing the capacity or resistance is to increase the time required for the control to change, when compensating for changes in the fork frequency. Accordingly the values chosen must be a compromise. If we assume that the control is capable of giving a maximum change in frequency of .1%, and we allow a fluctuation in this control of 5% each half second, this will cause a fluctuation in frequency of 5%, of .1%, or .005%. However, the inertia of the fork prevents it from following such a rapid fluctuation in damping current and hence the actual change in fork frequency is very much smaller than just indicated.

The fact that non-cumulative fluctuations in the control as great as 5% have only a negligible effect on the fork frequency is an important point. Such fluctuations are likely to arise through hunting in the synchronous motor, irregularities in the time of operation of the relay, etc., and since their effects average one another out, there is no danger of their being transferred to the fork.

The ratio of the charging resistance to the discharge or grid leak resistance is not a governing factor, except that the charging resistance must be less than the discharge resistance. The phase position of the fork to the clock under normal conditions is also governed by the relative values of these resistances. For the present circuit r_1 has a resistance one-half that of r_2, and these resistances and the condenser are of such values that it takes approximately 15 minutes for the fork to come into the correct phase relation with the clock when started under the most unfavorable conditions.

While this method of control will hold the fork frequency for an indefinite period in synchronism with the clock, it is possible that the phase relation of the clock to the fork may change. This change may be periodic, that is, it may take the form of an oscillation about the mean phase position, or there may be a gradual change due to changes in the various constants of the control occurring over comparatively long intervals. For instance, any change in the ratio $\frac{r_2}{r_1}$, such as might occur with temperature, will change the phase relation between the fork and the clock.

Chronograph records show that there are no phase changes greater than one cycle of the fork over periods as large as 8 hours. To determine the possibility of hunting, that is, of oscillation of the fork frequency around its mean value, the phase relation was actually dis-

turbed and a chronograph record taken of the readjustment. This will give the period of the oscillation, if any, and the amount of damping.

Fig. 3 shows one of these records. The chronograph was connected in the circuit as shown in Fig. 1 and a record was taken over a period of about 20 minutes after starting the fork. This record shows the length of time in each half second that the control relay was operated. At starting this period is about .11 second. After about 8 minutes it becomes a maximum equal to .2 second and there is no appreciable change over the next 5 minutes, showing a permanent condition has been reached. Accordingly we may conclude from this record that any oscillation about the mean value of the control

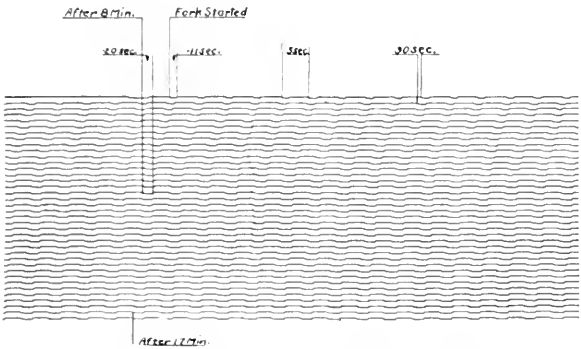


Fig. 3—Chronograph Record of Fork After First Starting

is almost if not quite critically damped. Several other records taken with even greater phase displacements bear out this conclusion. This practically precludes hunting after the phase angle has been once adjusted.

Accuracy of the Clock-Controlled Fork

The accuracy of the fork has been checked in two ways For long periods of time, chronograph records have been taken at intervals over a period of 8 hours and the maximum variation of the fork from the clock in this period has been found to be less than .02 second, or one cycle. Smaller periods of time cannot be measured accurately on the chronograph used. If we are dealing with periods of time of

more than 15 minutes, this gives an accuracy as high as one part in 50,000.

For small time intervals, an entirely different method for measuring the constancy of the fork must be used. Two methods are available. We may either compare the high harmonics of the fork directly with some high frequency which can be held extremely constant over short periods of time, and observe fluctuations in the relative values of these frequencies, or we may compare the fundamental frequency of the fork with a high frequency by some means which will enable us to measure the divergence from an exact integral multiple relation in terms of the higher frequency.

To explain in more detail, we may pick out the one hundredth harmonic of the fork by means of a filter and amplifier and compare it with a 5000-cycle frequency obtained from a constant frequency oscillator by some method of detection which will allow us to count the difference in cycles. By this means we may observe variations in the relative rate of the fork and the oscillator to an accuracy of about one-tenth of a cycle over a period of a few seconds, and this gives us a comparison to an accuracy of 1 part in 50,000. The principal objection to this method is the difficulty involved in separating the higher harmonics of any alternating current wave obtained from the fork. For instance, the separation of the hundredth harmonic from those immediately above and below it would require a circuit so selective that it would probably be very difficult to construct and cumbersome to operate.

If we had means to determine when some high frequency such as 5000 cycles was an exact multiple of the 50 cycles and to measure the difference in terms of the 5000-cycle wave, we would be able to obtain the same results, and avoid the above difficulty.

A device which will allow us to do this is the low voltage cathode ray tube developed by Johnson[3]. The two frequencies to be compared are connected to the two pairs of plates of the tube and the combination of the two deflections causes the luminous spot to trace out a path which repeats itself indefinitely if one frequency is an exact integral multiple of the other, and a stationary figure is produced. In this way any frequency which is a multiple of the fundamental 50 cycles may be accurately determined. As the method of comparison is an electrostatic one practically no power is used.

For the type of tube used, a deflection of about 1 centimeter is obtained for a potential difference of 10 volts between plates, and

[3] J. B. Johnson, *Bell System Technical Journal* Nov. 1922, "A Low Voltage Cathode Ray Oscillograph."

frequencies having ratios as high as 100 to 1 may be readily compared. For ratios of the order of 100 to 1 the lower frequency is preferably stepped up to a high voltage to give an equivalent deflection of as much as 25 centimeters, thus giving a spacing between cycles for the high frequency of approximately 0.5 centimeter. Of course, the whole 25 centimeter deflection is not shown on the screen but this is unnecessary. The value of the ratio cannot be at once determined by this means, there being no appreciable difference between the figure for a ratio of 100 to 1 and 99 to 1, but this ratio may be readily determined by comparing each frequency separately with an intermediate frequency such as 500 cycles.

Having determined the ratio between the high and low frequencies, it is possible, by drawing a reference line across the screen, to determine whether or not they are keeping step with one another. Thus for a comparison of 50 cycles against 5000 cycles, if we get a motion of 2 waves in 10 seconds, this represents a deviation from exact synchronism of 2 parts in 50,000.

Comparisons made in this way between the 50-cycle fork and a vacuum tube oscillator giving a constant frequency of 5000 cycles show no deviation in the mean period of the fork greater than 1 part in 50,000 for observations extending over several minutes. If deviations greater than this were observed, they might equally be ascribed to the auxiliary oscillator but the fact that they do not occur means either that the fork is constant to better than 1 part in 50,000 or that both frequencies vary in exactly the same way which is very improbable.

The above method of comparison does not require a sine wave of current from the fork. In fact it has been found advantageous to have a somewhat distorted wave since an unsymmetrical figure on the luminous screen of the tube is more easily observed. This is due to the fact that one-half of the figure moves across the screen in one direction while the other half moves in the opposite direction. In order not to confuse one half with the other, it is highly desirable that they be dissimilar in shape and this is accomplished by using a distorted wave as the lower frequency. Sufficient distortion is secured by mounting an ordinary telephone receiver in close proximity to one prong of the fork as shown in Fig. 1 and amplifying the e.m.f. thus obtained as much as necessary to obtain the desired voltage.

By means of the simple control system described above, it has been possible to obtain a fundamental frequency so free from fluctuations as to be constant over short or long periods of time to approximately one part in 100,000.

Accuracy of the Clock

So far we have not considered the possibility of error in the clock as a factor. Of course, the fork cannot keep better time than the clock which controls it.

The clock used at present was made by L. Leroy and Co., Paris, electrically driven and beating half seconds. The drive consists of an electric circuit including a single primary cell mounted in the clock, a driving coil and a contact which is closed by the escapement wheel for approximately .1 second in each second. Attached to the lower end of the pendulum is a steel bar which moves into the driving coil as the pendulum oscillates. The electrical impulse is so timed that the driving coil gives the pendulum a slight pull as it is entering the coil. This impulse is sufficient to keep the pendulum oscillating. An additional contact on the clock is used to furnish an electrical impulse for timing purposes.

Time records of the clock have been kept over a period of several months and the rate has been found to be constant to about one-half second a day, which is better than 1 part in 150,000. Since this accuracy is not very much greater than the precision with which the fork keeps in step, any further accuracy will require refinements in the clock itself. With this object in view, an investigation was made of the possibility of obtaining greater accuracy from the existing clock.

Errors are of two kinds. First, if the timing contact is obtained by the operation of the escapement wheel, there may be a cyclic variation in the length of time between successive impulses extending over one revolution of the wheel, (1 minute) even though the pendulum keeps perfect time. This has been found to be the case in some of the best clocks in the country. This error can be overcome by taking the contact direct from the pendulum. The contact we are using at the present time is of this type obtained from the pendulum by means of a photo-electric cell.

The optical system is shown on Fig. 4. Light from the source A is concentrated on the mirror, which in turn reflects it on to the photo-electric cell. When the pendulum passes through the center of its stroke, it momentarily cuts off this beam of light. This causes a large increase in the resistance of the photo-electric cell, the change taking place almost instantaneously.

Referring to the diagram of connections on Fig. 1, the potential of battery B is divided almost equally between the photo-electric cell and the grid of the tube if the grid leak is made approximately equal to the resistance of the cell when exposed to the light. This

gives a negative potential to the grid sufficient to cut off all space current, and the relay R_2 remains unoperated. When the pendulum cuts off the light to the photo-electric cell, the resistance of the cell rises immediately and the grid voltage drops to a very small value. Enough space current will pass now to operate the relay R_2 and a

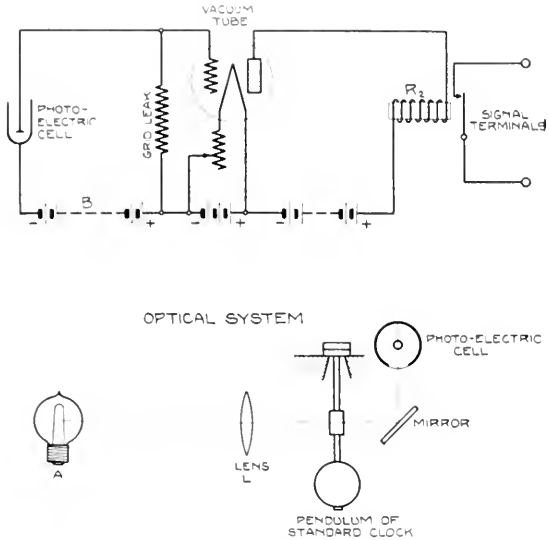


Fig. 4—Circuit for Obtaining Electrical Impulses from Standard Clock Using Photo-electric Cell

signal is transmitted of the same duration as the time the light is cut off the cell by the pendulum. There is no appreciable time lag in the photo-electric cell or vacuum tube.

The principal requirement in setting up this circuit is to obtain a vacuum tube having a resistance between filament and grid including wiring, which is under all conditions considerably greater than the minimum resistance of the photo-electric cell. If this resistance drops much lower, the circuit becomes inoperative even though no additional grid leak is used.

The only irregularity introduced in this system is in the operation of the relay, and as this is a fast operating relay this error will be

less than the accidental irregularities in a contact obtained from the escapement wheel even excluding errors due to eccentricity.

This method of obtaining an electrical impulse from a clock is of great value as it may be applied to practically any clock which may not have any other method of producing impulses.

The second type of error is due to variations in the rate of the clock. Two fundamental requirements in the design of an accurate clock are that the impulse delivered to the pendulum be symmetrical about the mid-point of its swing and be not subject to irregularities in magnitude or duration, and that the pendulum be free at all other parts of its swing. These requirements are fairly well met in the

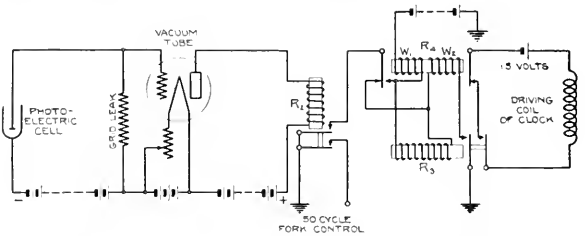


Fig. 5—Circuit of Photo-electric Cell Drive of Standard Clock

present clock. However, the magnitude of the impulse depends on the constancy of the voltage of the driving cell which is a single primary cell of rather small size, and the duration of the impulse may be somewhat variable due to the possible eccentricity of the escapement wheel and due to the method of operation. The pendulum too, is not entirely free from constraint at any part of its swing. These errors may all be avoided or at least considerably reduced by the use of the impulse obtained from a photo-electric cell to drive the clock and by the use of a more constant source of primary voltage.

The use of this type of drive has accordingly been investigated in connection with this clock. It is obvious, since the driving impulse is one of attraction between the coil and the bar carried by the pendulum, that it must be exerted only once per second, that is, when the pendulum is entering the driving coil and not when it is returning. The circuit used is shown on Fig. 5 and operates as follows:

When the relay R_2 operates the first time in the second, it closes the circuit through the winding W_1 of the relay R_1 and through relay R_3. This operates the relay R_3 and closes the circuit through the driving coil of the clock. The current through the one winding of

R_4 is not sufficient to operate it. As soon as the relay R_2 releases, current will pass through all the windings on both relays which in turn closes the relay R_4. This opens the circuit through the driving coil of the clock. The impulse given to the pendulum is, therefore, the duration of the operation of the relay R_2, or the time during which the light is cut off the photo-electric cell during the swing of the pendulum to the left. When the pendulum swings to the right and the relay R_2 operates, R_3 is short-circuited and releases, R_4 being held up by winding W_4. When R_2 releases, it releases R_4 bringing the circuit back to normal. Since the circuit through the driving coil of the clock is closed only when the relay R_3 is closed, and the relay R_4 is released, there is only one impulse per second given to the pendulum.

During a period of operation by this method covering several days the clock gave as satisfactory performance as with the mechanical drive, but while the present gear train is connected to it, no appreciably better performance can be obtained than at present, and accordingly it is proposed to carry out further work along this line with an experimental pendulum having no mechanical connections. By using a good compensated pendulum and mounting it suitably in a constant temperature hermetically sealed case, it appears probable that a photo-electric cell drive would produce a more constant rate of oscillation than the best clocks of existing types. The advantage of this type of drive over other types is the fact that the pendulum is absolutely free from all mechanical constraint at all parts of its swing. The problem of supplying an uninterrupted current for the light and power could readily be solved by the use of duplicate apparatus.

The general method outlined in this paper for synchronizing a fork with a clock has a very wide field of usefulness, and is not limited to the particular application described. For instance, in place of the clock we may substitute another fork and distributor, and we are thus enabled to hold 2 forks with their distributers in exact synchronism by means of an impulse transmitted at a constant time interval of about once every half-second.

By substituting the field coils of a motor for the damping winding on the fork, we are able to hold the speed of the motor in synchronism with the clock, the only requirement being a step down gear on the motor to furnish the desired contact.

The general principle involved is not dependent on the use of a vacuum tube, and if other means of control based on this principle be adopted, very large powers may be controlled in the same way.

Some Contemporary Advances in Physics—II

By KARL K. DARROW

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

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

The present installment, which is Number II, is devoted very largely to the subject of atomic structure.—EDITOR.

WE know quite definitely that an atom consists of a massive positively-charged nucleus with a certain number of electrons in its vicinity; but of the arrangement of these electrons in the strict geometrical sense we know very little—indeed, we do not certainly know even whether they are in motion or not. Apparently there are many possible arrangements for each kind of atom; one of these is a permanent arrangement, in the sense that when once established it is not changed so long as the atom is not disturbed from outside; the others are transient. In addition to the arrangements of the electrons in the neutral atom, there are the arrangements of the remaining electrons when one or more of the normal quota are lacking. When an atom changes over from one of these arrangements to another, it must take in or give out a definite quantity of energy. Another way of saying this same thing is that to each distinct arrangement of the electrons there corresponds a distinct value of the energy of the atom. These values of the energy of the atom are directly or indirectly measured, often with great precision; they are the data of experiment. The very precise statements, or at all events very definite statements, which are frequently made about the "structure" of the atom, usually refer only to these energy-values and the relations between them.

The simplest question that can be asked about the arrangement of the electrons is, whether they all occupy identical positions—being, for example, evenly distributed over the surface of a sphere or the circumference of a circle, with the nucleus at its centre. If this is true, the same amount of energy will be required to remove any

electron from the atom as to remove any other. In the extreme opposite case there would be as many different amounts of energy required to remove an electron from the atom as there were electrons. Now, when radiation of a definite frequency ν falls upon a group of atoms, any particular atom will either ignore the radiation, or else will absorb a definite quantity of energy $h\nu$ from it. (The letter h, as usual, denotes Planck's constant, 6.56×10^{-27} ergs-seconds.) It

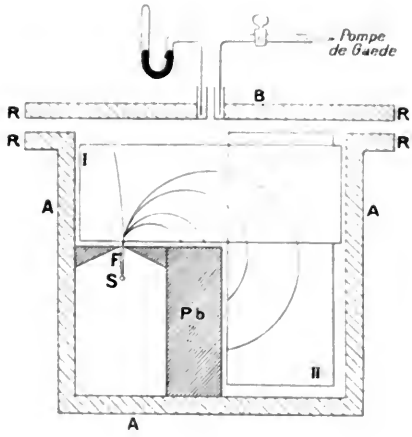


Fig. 1

follows that if an electron is extracted from an atom by this radiation and the work W required to extract it is not exactly as great as the amount $h\nu$, the difference will be turned over to the electron as kinetic energy, and the speed v with which it departs from the atom will be given by the equation

$$\frac{1}{2} mv^2 = h\nu - W$$

and W can be determined by measuring v. We can conveniently refer to W as an "extraction-energy" or "extraction-potential." If all the electrons occupy identical positions, W will be the same for all, and the emerging electrons will all have the same speed. If they occupy various positions or "levels" as is more commonly said, there

will be as many different electron-speeds represented in the emerging electron-stream as there are levels,[1] and from these speeds the extraction-energies characterizing (or indeed defining) the levels can be deduced.

The apparatus in which the test is made is of the type shown in Fig. 1. At S there is a long narrow rod or tube of the material being tested, irradiated by X-rays proceeding from a source at the left. A magnetic field, directed normally to the plane of the paper, sweeps

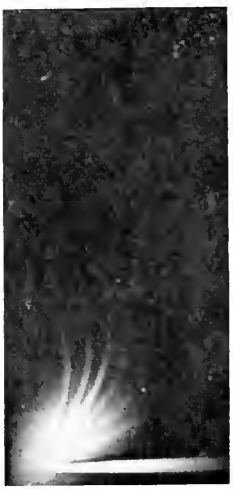


Fig. 2

the emerging electrons around in circular arcs, some of which pass through the slit; a few such arcs are sketched. The slower the electron, the more highly curved the path in which it travels; and the speed of the electron can be deduced from the curvature of the path. In Fig. 2 electron-paths of this type are reproduced from a photographic film, which was laid parallel to the plane of the paper, in the position of the rectangle marked I in Fig. 1. Fig. 3 shows arcs which appeared on a film laid in the position of the rectangle marked II.

[1] Of course there may be reasons why electrons in particular positions cannot often or at all be extracted by radiation, even though there is plenty of energy available.

These distinctly-separated arcs show that the emerging electrons fall into several distinct groups, each characterized by a particular speed. In Fig. 4 we see the traces of the electrons on films laid normally to

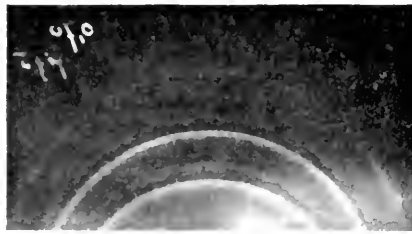


Fig. 3

Fig. 4

the plane of the paper, along the top of the block marked "Pb" in Fig. 1. The appearance of the films at once suggests a line-spectrum. The lines, indeed, are the signatures of special electron-speeds instead of special radiation-frequencies; but these two quantities, being interconvertible, are not so profoundly different in their nature as

used to be supposed. Imitating de Broglie's term "spectres corpusculaires" we may call these "electronic spectra." But it must be remembered that they depend not only on the properties of the atom, but on the incident radiation as well.

Maurice de Broglie has undertaken an extensive study of these electronic spectra. His most recent apparatus, similar in general to the arrangement illustrated in Fig. 1 (with the photographic plate laid normally to the plane of the arcs) is improved in various respects and enlarged to permit of using a plate 24 cm. wide and electron-paths of 26.cm. radius. Unfortunately, the ideal condition of atoms irradiated by radiation of a single frequency, is unattainable. This is not merely because actual X-ray sources emit very mixed radiations intense at several distinct frequencies and perceptible at every frequency over a wide range. This difficulty could be partly remedied by appropriate filters. There is another difficulty and an inevitable one; the atoms from which electrons are extracted by the radiation promptly emit radiation of new frequencies, which extract other electrons themselves. In the language of the opening paragraph, the arrangement of electrons which results when an electron is extracted is not a permanent one; the remaining electrons redispose themselves in one arrangement after another, eventually arriving at the permanent one; to each successive arrangement corresponds a new and lower value of the energy of the atom, and the energy-differences ΔE are successively sent out in radiations of frequencies $\Delta E\ h$. Thus there are several frequencies at work extracting electrons from the atoms; and in the electronic spectrum, each level is represented by as many lines as there are frequencies.

The uppermost spectrum of Fig. 5 is sketched by de Broglie from photographs made with the electrons emitted by silver atoms irradiated with the characteristic X-rays of tungsten.[2] The electron-speeds corresponding to the lines increase from left to right. There are four of these tungsten rays, two forming the $K\alpha$ doublet, while the other two, known as $K\beta$ and $K\delta$, have higher frequencies. The four lines marked 4 and 5 in the electronic spectrum are made by electrons extracted by these four radiations from a single level. This is the K-level, the deepest or innermost level in the silver atom, the electrons removed from it having lost more energy during the removal, than any others observed, about $3.46 \cdot 10^{-8}$ ergs apiece. The two following doublets, marked 6 and 7, are made by electrons extracted by the $K\alpha$ frequencies from two distinct levels of the silver atom,

[2] Some photographs may be seen in the *Journal de Physique*, volume 2 of 1921. They were taken before the latest improvements were made in the apparatus, and do not show so much detail as the sketches; or perhaps the reproductions are imperfect.

known as the L and M levels respectively; the electrons from them have more energy left over after escaping. Line S is due to $K\beta$ extracting electrons from the L-level. The electrons ejected from the M-level by $K\beta$, and those ejected from the L and the M levels by $K\delta$, are presumably moving too rapidly to be received on the plate. At the other end of the spectrum the three lines 1, 2, 3 are due to electrons

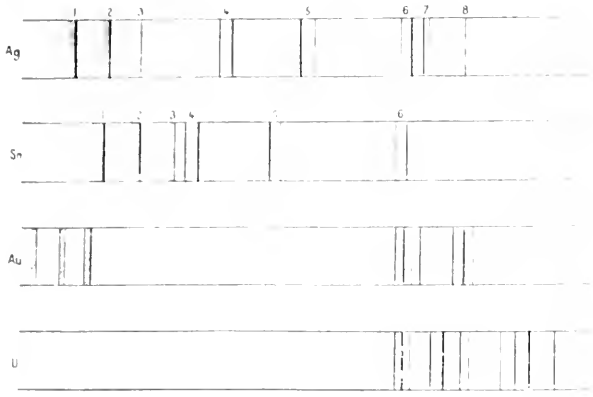


Fig. 5

expelled from the L and the M levels by two of the secondary X-ray frequencies proceeding from silver atoms: the $K\alpha$-doublet (not separated) and the $K\beta$-line of silver.[3] Just below this spectrum, we see the electronic spectrum of tin, in which the lines due to the primary X-rays from the tungsten are arranged like the corresponding lines in the silver spectrum, but displaced towards lower energies, since the levels in the tin atoms are different from those in the silver atom; while the lines due to the secondary X-rays are also repeated from the silver spectrum but with an opposite displacement, for in these cases both the levels from which the electrons are taken and the energies available for taking them out have been changed. Next come the spectra of gold and uranium. Each of these elements has more electrons per atom than the previous two (uranium has more

[3] From the nature of the rearrangements resulting in the $K\alpha$ and $K\beta$ radiations, it follows that the electrons extracted by the former from the M level have the same speeds, very nearly, as those excited by the latter from the L level, the two frequencies acting on the two levels produce three separable lines.

electrons, ninety-two, than any other element). The complexity of the spectra results from this richness of electrons, but the electrons extracted from the L and M levels of gold by its own radiations can be identified.

It is not necessary to provide an X-ray tube to supply the primary radiation; this can be supplied from the nuclei of radioactive atoms mingled with the atoms being tested, or, by examining radioactive substances, we can discover electronic spectra excited by radiations originating at the nuclei of the atoms themselves. Actually these were the earliest electronic spectra discovered; the first to be observed were photographed by von Baeyer, Hahn, and Meitner in 1910, years before the interpretation was made (the frequencies of the nuclear radiations were not then known). The figures 1, 2, 3 and 4, used to illustrate this article, are taken from a paper by J. Danysz, describing work performed in 1911 at the laboratory of Madame Curie in Paris, upon the electrons or beta-rays emerging from atoms of radium B and radium C. The grouping of these electrons, as we now know, results from their being extracted from the various levels by the several nuclear radiations and the inevitable secondary radiations which they produce in their own atoms. The large number of distinct groups (Rutherford and Robinson distinguished sixteen from radium B and forty-eight from radium C) is very likely due to several co-operating causes; there are several frequencies at work, the atoms have large numbers of electrons, and extractions probably occur exceptionally often where the radiations originate so close to the electrons. The earliest electronic spectra produced from non-radioactive atoms were excited by nuclear rays from radioactive substances, and the earliest rule discovered was that these spectra were very similar to the spectra of the radioactive atoms themselves; being indeed identical when the excited atoms are isotopes of the atoms which emit the exciting rays. In a complete account of this topic, many other names would be mentioned, notably those of C. D. Ellis and R. Whiddington.

A recently-published and relatively simple case is that of the radioactive atom, uranium X_1, of which the electronic spectrum is shown in Fig. 6 (from an article by Frl. Meitner). This displays three lines made by electrons of which the speeds indicate that they are extracted from the L, M and N levels of the atom by a single radiation, having itself the frequency of the natural $K\alpha$-radiation of the atom. This radiation was itself detected and identified by appropriate means. Faster electrons which were also observed, cannot have been derived from any such source; they probably came from

the nucleus, and some of them eject electrons from the K-level of the atom, thus producing the necessary condition for the $K\alpha$-radiation and all the others to be emitted. These electrons from the K-level would escape with too little energy to be registered in the apparatus. The question of the ultimate origin of these fastest electrons is, however, still under debate by the leading authorities on the subject.

Fig. 6

Imagine now that a beam of X-rays including all frequencies is directed against a thin sheet of metal atoms, and that the transmitted beam is dispersed into a spectrum projected against a photographic plate in the usual manner. Rays of frequency ν can extract electrons from a particular level when $h\nu$ exceeds the value of W for that level, but not otherwise. Advancing along the spectrum in the direction of increasing frequencies, we should expect to find a sudden sharp weakening of the transmitted rays wherever the frequency becomes equal to one of the values W/h which characterize the various levels. Some of L. de Broglie's classical photographs are shown in Fig. 7 (borrowed from Millikan's book, "The Electron"). The second picture from the top represents two spectra of an X-ray beam transmitted through molybdenum, one spectrum stretching away to the right from the central dark band, the other to the left. The frequency decreases as the distance from the dark band increases. Coming inwards toward the band, we see that the plate very suddenly becomes whiter at a certain critical frequency; this is the frequency at which $h\nu$ becomes equal to the W of the K-level. Similar spectra of beams transmitted through cadmium, antimony, barium and mercury are presented below the molybdenum spectrum; the corresponding absorption-edge is discerned in each, its frequency rising with the atomic number of the element.[4] This is by far the most delicate and accurate method of determining the various extraction-energies,

[4] The topmost picture shows the spectrum of the beam before it encounters the absorbing layer; the various strong lines in it, and the absorption-edges impressed upon it by the silver and bromine atoms in the photographic film, recur more or less clearly in the absorption-spectra, but have nothing to do with the atoms in the absorbing layer.

although by itself it merely shows that particular transformations of X-ray energy become possible at particular frequencies. The electronic spectra, although much less accurate for purposes of measurement, are needed to show how this absorbed X-ray energy is used.

These absorption-spectra show that the levels in the atom are much more numerous than the electronic spectra with their lower "resolving power" can reveal; for example, there are three L-levels and five

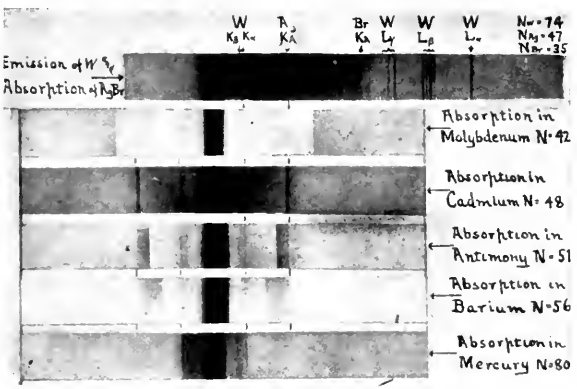


Fig. 7

M-levels. The five M-levels of thorium display themselves (not as clearly as might be desired) in Fig. 8, which consists of absorption-spectra photographed by P. A. Ross at Leland Stanford University. Each spectrum extends from low frequencies at the left to a maximum limiting-frequency at the right, the limit depending on the voltage applied to the X-ray tube and not on the properties of the absorbing atoms. As the limiting-frequency is increased by increasing the voltage, the absorption-edges resulting from electrons being extracted from the five M-levels successively appear. Along with each new absorption-edge there appear one or two new emission-lines, emitted by the thorium atoms during the rearrangements which follow upon the extraction of an electron.[5] These correspond

[5] Actually these lines were not emitted by the same atoms as absorbed the X-rays, but by thorium atoms in the target of the X-ray tube whence the primary X-rays came; the preliminary electron-extractions were performed in most cases by swift electrons. There is no reason to suppose that the agency effecting the electron-extraction has anything to do with the subsequent rearrangements of the atom.

to the secondary radiations of silver and tin, which we found to produce lines of their own in the electronic spectra of these elements. These relations between absorption-edges and emission-lines make

Fig. 8

it possible to use the X-ray emission-lines of atoms to identify and map out their levels.

This discussion of electronic spectra and X-ray absorptions has served to illustrate the remark made in the opening paragraph, that our knowledge about the various arrangements of the electrons

forming the atom consists mainly of data about their energy-values. We have a key to the arrangements themselves, and this is provided by the deflections of electrons as they pass through the atoms. An electron shot directly at an atom will be deflected by the combined actions of the nucleus and the atom-electrons; and by postulating a particular arrangement of the electrons we could, in principle at least, calculate the deflection. This may be likened to the performance of an astronomer who, observing a comet advancing into the solar system from outer space, calculates the path which it will follow through the system under the influence of the sun and the major planets, and the direction along which it will depart. The astronomer has the advantages of knowing exactly where the members of the solar system are, and of being able to follow individual comets. We do not know where the members of the electron-system are, and cannot shoot a single electron at an atom and discern its path.

The latter disadvantage is not as serious as it may seem. By projecting an enormous number of electrons in parallel directions against an atom or a layer of atoms, and measuring the fraction which are deviated through a given angle or range of angles, it is possible to test a particular atom-model. Assume that the atom possesses spherical symmetry; then the deflection suffered by an oncoming electron will depend only on a single variable, the minimum distance p from the centre of the atom to the line (extended) along which the electron approaches at first (before the deviation begins). Designating by ϕ the angle between the initial and final directions of motion of the electron (i.e., the amount of the deflection), we have

$$\phi = f(p) \qquad p = f^{-1}(\phi) \tag{1}$$

the function f depending on the particular atom-model. Suppose an enormous number N of electrons directed normally against a thin layer of metal atoms, in which Q atoms lie side by side. The number of electrons which will approach the layer along lines passing some atom-centre —any atom-centre— at distances greater than a given value p and less than a slightly greater given value $p+dp$, is

$$dN = NQ \cdot 2\pi p \cdot dp \tag{2}$$

This is likewise the number of electrons which will be deflected through angles lying between $\phi = f(p)$ and $\phi + d\phi = f(p+dp) = f(p) + (df/dp)dp$; which therefore may be written as

$$dN = NQ \cdot 2\pi p (dp/df) df = F(\phi) d\phi. \tag{4}$$

The expressions for p and $dp/d\phi$ are to be taken from equation (1). The function

$$F(\phi) = NQ \cdot 2\pi p(dp/d\phi) \tag{5}$$

represents the *distribution-in-angle* of the deflected electrons. If it is calculated for any particular atom-model and then determined by experiment, the comparison between calculation and data affords a test of the atom-model. An instructive comparison can be made even if the value of NQ is unknown, since the form of the F-*versus*-ϕ curve, as well as the absolute height of its ordinates, depends upon the atom-model.

For electrons or other charged particles of charge e and mass m, streaming with uniform speed U against a group of much more massive nuclei each bearing a charge E, the functions f and F assume the forms

$$f(p) = 2 \cdot \text{arc cot}\,(mU^2p/eE) \tag{6}$$

$$F(\phi) = NQ\pi(eE/mU^2)^2 \cot\,(\tfrac{1}{2}\phi)\,\text{cosec}^2(\tfrac{1}{2}\phi). \tag{7}$$

This case, insignificant as it may appear, suddenly assumed the greatest importance when, in 1913, Rutherford, Geiger and Marsden established that the distribution-in-angle of alpha-particles (particles of twice the charge and about 7,500 times the mass of an electron) deflected by metal atoms is of precisely the form (7). This means that around each atom-nucleus there is an empty space so wide that full-speed alpha-particles passing close enough to a nucleus to be deflected through 5° or more, undergo almost their entire deflection within it; hence, most or all of the electrons surrounding the nucleus must lie beyond this vacant central region. From the data of these classical experiments, Rutherford and his collaborators deduced that the radius of the empty region encircling the gold nucleus is at least 36×10^{-12} cm. After the war, the problem was again taken up in Rutherford's laboratory in Cambridge. J. Chadwick gave 14×10^{-12} cm. as a minimum value for the radius of the vacant space around the platinum nucleus. Last year P. M. S. Blackett made a statistical study of the deflections of comparatively slow alpha-particles, using the C. T. R. Wilson expansion-method, which was described in the last issue of this Journal. Paths of some of these deflected particles are shown in Fig. 9. By using these slow-moving particles, which begin to turn in their courses while still much farther away from the nucleus than the minimum distance at which fast alpha-particles begin to respond to its repulsion, Blackett was able to search farther

out for the outer boundary of the empty space. The deflecting atoms were atoms of argon, each consisting of eighteen electrons surrounding a nucleus; atoms of oxygen with eight electrons apiece, and atoms of nitrogen with seven (the latter two kinds of atoms not being discriminated in the study of the data). Blackett concluded that the

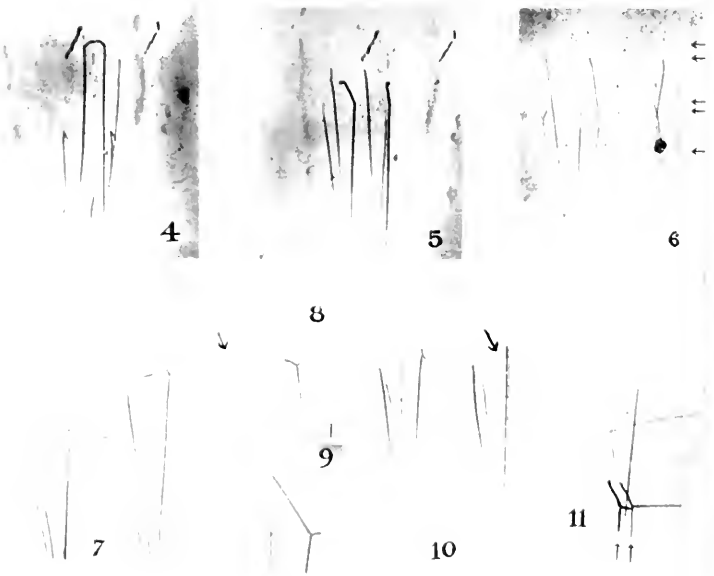


Fig. 9

empty space in the argon atom extends out at least to a distance of 40×10^{-8} cm. from the nucleus in the argon atom, and to a distance of at least 5×10^{-8} cm. in the nitrogen and oxygen atoms.

We now pass to the case of electrons deflected by atoms. Since the electron is so very much lighter than the alpha-particle, and yet is half as strongly charged, it will be much more seriously deflected by a nucleus than an alpha-particle, approaching along the same line with the same speed, would be. This contrast is very strikingly illustrated by two results published last summer. Harkins and Ryan, photographing the paths of eighty thousand alpha-particles

through air, found only three instances of deflections exceeding 90°. C. T. R. Wilson, photographing the paths of 503 fast electrons through air, found forty-four instances of deflections exceeding 90°. While, in general, it would be hardly fair to make such comparisons without allowing for the relative energies of the two kinds of particles, the difference in order of magnitude is so great that we may accept it as typical.

Moreover, the electron will be deflected by the atom-electrons as well as by the nucleus, and will not disarrange the atom-electrons so badly on its way through the atom-system. These deflections

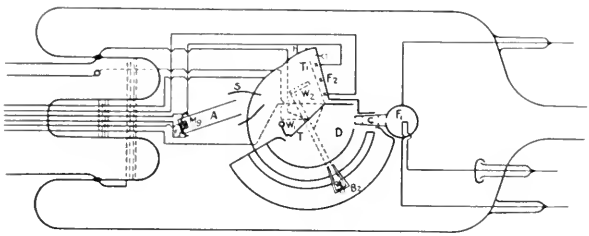


Fig. 10

will be superposed upon the deflection produced by the nucleus, and will modify the distribution-in-angle function F from form (7) into some other form. Such modifications have been suspected by several investigators; for example, by Crowther and Schonland in their study of the deflections of very fast electrons by metal atoms. It has been argued by Wentzel, however, that the distribution-in-angle function observed in their experiments departed from the form (7) not because the atom-electrons were interfering with the fast electrons, but because some of the deflected electrons had been deviated by several atom-nuclei in succession.

C. Davisson and C. H. Kunsman, in the laboratories of the Western Electric Company, made the first definite attempt to produce electron-deflections under conditions in which the distribution-in-angle function would disclose the influence of the atom-electrons. To do this it was desirable to use, not the fast electrons from radioactive atoms which previous experimenters had employed, but slow electrons of controllable speed. A diagram of their apparatus is shown in Fig. 10 and a photograph in Fig. 11. The electrons proceed from a hot filament at F_1, strike the metal target at T, and are deflected through various angles; the shielded collector B_1, swinging from one angle to

another, successively receives the electrons deflected through the various angles. The electrons depart from F_1 with very low speeds and receive the speed U through acceleration by a voltage applied between F_1 and the cylinder surrounding F_2; thereafter they move with constant speed in an equipotential region, through the various

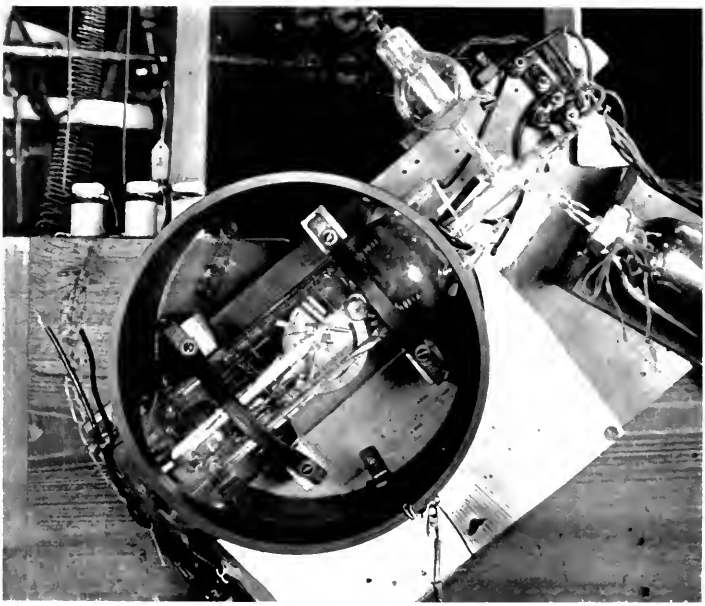


Fig. 11

slits shown in the diagram around C, and against the target. Among the electrons which emerge from the target, there are some which have been deflected by individual atoms in the way we have been describing, but very many more which have either undergone several deflections in succession or else were not in the incident beam but have been dislodged from their places in the target metal by the primary electrons. If these latter were allowed to reach the collector, the distribution-in-angle function of the once-deflected electrons would be blurred and concealed by the unwanted electrons. As,

however, they have not so much energy as the primary or the once-deflected electrons, they can be kept away from the collector by lowering its potential to a value such that only such electrons as have, say, 90% of the energy of the primary electrons can reach it. Thus the filament may be at potential zero, the target at 500 volts; if the collector is also at 500 volts, the distribution-in-angle function of the

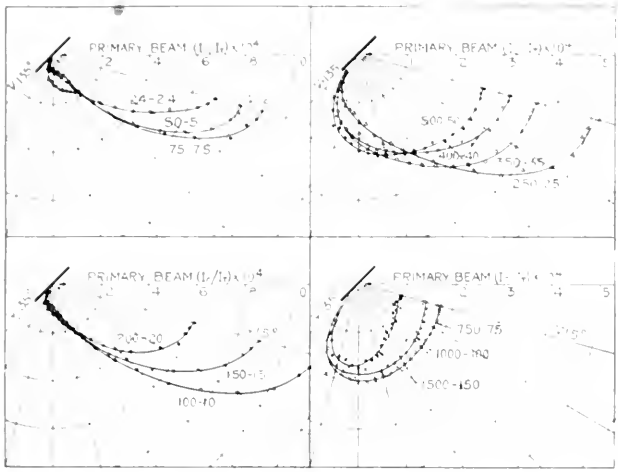


Fig. 12

electrons it receives has nothing in common with the function F characterizing the once-deflected electrons; but when the collector is lowered to 50 volts, the distribution-in-angle function which it records assumes a new and characteristic form.

Some of these angular distributions are shown in Fig. 12 (for magnesium) and Fig. 13 (for platinum). The latter curves were obtained first, with a platinum target; then the target was overlaid with a thin film of magnesium, formed by sublimation without opening or altering the tube, and the sharply-contrasted curves of Fig. 12 replaced the others. The distribution-in-angle of the engineering electrons is plotted, naturally, in polar coordinates; the direction $\phi = 0°$, i.e., the direction of motion of the primary electrons, is indi-

cated by the arrow and the lettering. Such a symbol as "100 10" indicates that the corresponding curve was taken down with the target at 100 volts and the collector at 10 volts (the filament always being at zero potential). The reason for this has been explained above; the family of curves in Fig. 12 illustrates the point.

These are examples of the curves from which the arrangement of the atom-electrons is to be inferred. The sinuous and serrated curves for platinum, entirely different from the smoothly rounded curves

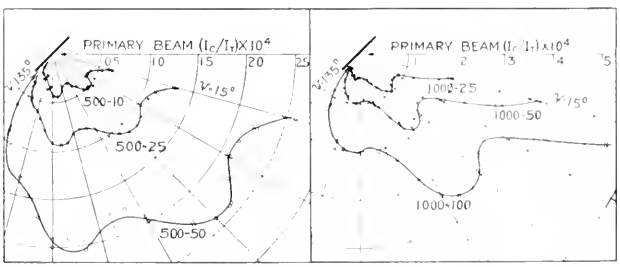


Fig. 13

derived from equation (7), surely owe their shape to the numerous levels among which, as was shown in the foregoing pages, the electrons of massive and electron-rich atoms are distributed; the platinum atom, with its seventy-eight electrons, ranks among the most complicated of all. The magnesium atoms, with their thirteen electrons apiece, are simpler and yield curves which are simpler, but not of the type of equation (7).

To interpret these curves Davisson has calculated the distribution-in-angle function for electrons deflected by an idealized "limited-field" atom-model, in which there is a concentrated charge $+E$ at the centre and a charge $-E$ uniformly spread over a spherical surface of radius R. This uniformly-charged sphere is a sort of first-approximation substitute for a spherical surface on which several electrons are arranged. It is not implied that the magnesium atom has all its electrons at the same distance from the nucleus, which would be most improbable, as its X-ray spectrum shows at least two distinct levels; we can suppose that n out of the 12 electrons are so close to the nucleus that together with it, they practically

form a single point-charge $(12-n)e$, and the remaining $(12-n)$ electrons lie on the spherical surface, which encloses the "empty space" mentioned above. The functions f and F assume the forms

$$f(p) = 2 \text{ arc cot } \frac{p(2\mu - 1)}{\sqrt{R^2 - p^2}} \tag{6a}$$

$$F(\varphi) = N Q \pi \cot(\tfrac{1}{2}\varphi) \operatorname{cosec}^2(\tfrac{1}{2}\varphi) \frac{(2\mu-1)^2 R^2}{[\cot^2(\tfrac{1}{2}\varphi) + (2\mu - 1)^2]^2} \tag{7a}$$

in which $\mu = \frac{1}{2} m U^2 R / eE$; the symbols have the same meanings as in (6) and (7), with which these equations become identical if R is made infinite.

This "limited-field" distribution-function has some odd characteristics. At very high speeds large deflections naturally are rare, but as the speed is lowered they become relatively more frequent; the 1,500-volt, 1,000-volt and 750-volt curves for magnesium illustrate this. This tendency gains rapidly as U is decreased; at a certain critical value, given by $\mu = 1$, the deflections are uniformly distributed in all directions[6]; at a lower critical value, given by $\mu = \frac{1}{2}$, all the electrons are turned through 180° and return on their tracks. As the speed is still further decreased, the condition of uniformly-distributed deflections is again approached, and we have the extraordinary feature of the average deflection decreasing as the energy of the electrons goes down.[7] In the family of curves for magnesium there appears very clearly an intermediate velocity at which 180° deflections are peculiarly frequent; the curves spread outward in the direction $\varphi = 180°$ whence the primary electrons come, as the energy of the primaries rises from 21 to 75 volts, and retract themselves again as the energy rises beyond 100 volts. This is a particularly important feature of the curves.

To make an adequate test of the new expression for F, it is necessary to apply certain corrections to the curves presented, particularly a correction required because the distance travelled by the deflected electrons within the target metal varies with φ, so that the percentage which goes astray, owing to loss of speed or otherwise, varies similarly. A curve exempt from this correction can, however, be ob-

[6] Meaning that the number deflected per unit solid angle is independent of φ, which means that the distribution-in-angle function is of the form $const. \sin \varphi$.

[7] It may be recalled from the last number of this Journal, page 110, that H. A. Wilson used this property as an explanation of the anomalous variations of electron-mean-free-paths with speed in various gases.

tained in a certain manner.[8] On studying this curve, it is found that the critical speed at which 180° deflections are most frequent is too low. This indicates that an incident electron approaching an atom is accelerated toward it, by virtue of the total charge of the electrons on the spherical shell not quite compensating the nuclear charge; the speed U which figures in the equations is therefore greater than the measured speed with which the electrons are fired at the target. (This interpretation also serves to explain the lobe observed on the lowest-speed curves for magnesium, and suggests the reason for the lobes of the curves for platinum.)

The curves are satisfactorily explained, if we build the magnesium atom in this manner: a nucleus of charge $12e$, two electrons so near it that the central charge is effectually $10e$, and a spherical shell of six electrons with a radius of $1.28 \cdot 10^{-9}$ cm.; the other four electrons much further out, perhaps dispersed and wandering through the metal. The only arbitrary assumption made is that about the two deep-seated electrons; the radius R of the shell and the number of electrons upon it are prescribed by the curves, once that assumption is made. If we assume three deep-seated electrons, R becomes $1.15 \cdot 10^{-9}$ cm. and the number of electrons in the shell drops to five. The shell must be the L-level, and the deep-seated electrons constitute the K-level.[9]

The energy required to remove the loosest or outermost electrons of the atom is generally determined, as is well enough known, by smiting the atom with an electron instead of with one of the radiation quanta used in extracting the inner electrons.[10] Usually the quantity measured is simply the energy which the striking electron must have, in order to convert the atom or molecule into a positively-charged ion; the negative charge removed from the atom is assumed without proof to be a single electron. On the other hand, J. J. Thomson

[8] Imagine an electron incident at angle θ on the target surface, and deflected through angle ϕ (in the plane of incidence) by an atom which it meets after penetrating a distance d in a straight line. If it continues in a straight line from the point of deflection until it emerges, it travels a distance $x = d\,(1+\cos\theta \cdot \sec(\psi-\theta))$, where $\psi = \pi - \theta$. This distance x will be the same for any two values ψ_1 and ψ_2 of ψ, such that $\psi_1 + \psi_2 = 2\theta$. Insofar as the number of deflected electrons emerging with speed sufficient to reach the collector depends on x, it will be the same for both values of ψ. The curve representing the ratio of the number of electrons reaching the collector, for two such angles, plotted versus U, is exempt from this correction, and can be directly compared with a theoretical curve.

[9] Or we could assume that there were no deep-seated electrons, and give seven electrons and a radius $1.54 \cdot 10^{-9}$ cm. to the shell; but then we should have nothing to serve as a K-level.

[10] Generally the frequency required to extract the outermost electron with a quantum lies in the most inconvenient region of the spectrum for practical work.

and many others have measured the charges [1] of ionized atoms in discharge-tubes, and found them sometimes single and sometimes multiple electron-charges, but have not measured the minimum energy required to produce a particular kind of ionization. H. D. Smyth, at Princeton and Cavendish, was the first to combine both methods; he ionized atoms by electron-impacts in a tube designed for determining ionization-potentials in the accepted manner, and after further accelerating the ions drew them through a channel into a second tube where they were deflected in a magnetic field so that their charges could be measured. The difficulty to be surmounted is that in the first tube the pressure of the gas must be high enough to yield a satisfactory number of ions, and in the second tube it must be low enough not to interfere with the arcs described by the ions in the magnetic field. At first he sent a beam of mercury vapor rushing transversely across his first tube from a boiler into a liquid-air trap; by first sending the atoms down a long tube with a system of diaphragms and so stopping the obliquely-moving ones, he was able to prevent atoms from straying out of the beam in the critical zone. Later he attacked a more difficult case, that of nitrogen; the gas was continuously fed into the first tube and a powerful pump drew it out before it could diffuse seriously into the second tube.

While it is interesting to have direct confirmation that the first and easiest ionization is the extraction of a single electron, Smyth's most important results refer to the later ionizations. Mercury atoms that had lost two electrons appeared in the second tube when the bombarding potential attained 19 volts, nine volts more than the first ionizing-potential; at a much higher voltage, triply-charged atoms were detected, or at least suspected. In nitrogen, the earliest ionization, at about 16 volts, does not involve dissociation, but at a potential 8 volts higher, a doubly-ionized single nitrogen atom makes its appearance, and a little further along, Smyth detects an ion which may be a singly-ionized nitrogen atom or a doubly ionized molecule (the two possibilities cannot be discriminated by this method, but the second seems improbable). Valuable knowledge about the relations between ionization and dissociation between, that is, the removal of an electron from a molecule, and the breaking of the bonds that hold the atoms of the molecule together may be expected from experiments of this type.

Something more is to be said on two of the topics of the last article in this series. A. H. Compton's discovery that scattered X-rays consist of two distinct radiations, one with the frequency of the

[1] Actually, the charge-mass ratios.

primary rays and the other with a slightly lower frequency, was mentioned in that article; he has since published an account of a series of measurements made, not on the wave-length but on the absorption-coefficient (in various substances) of the scattered rays, and finds it altered from that of the primary rays in the sense and more or less in the magnitude to be expected from the wave-length measurements of the lower-frequency rays. The largest alterations and the best agreements with theory are obtained with light atoms and high-frequency rays. In the frequency-range of the visible spectrum, the scattered ray of lowered frequency, sought for by P. A. Ross in light of the wave-length 5461A scattered by mercury vapor, is altogether lacking. The transparency of krypton and xenon atoms to slow electrons, discovered by Minkowski and Sponer, has been confirmed by Ramsauer with his original (and better) method. The transparency of argon atoms has also been verified by O. W. Richardson and R. N. Chaudhuri, by a method sufficiently different from the others to rank as an independent test.

REFERENCES

O. von Baeyer, O. Hahn, and L. Meitner: Phys. ZS, *11*, pp. 488-493 (1910); *12*, pp. 273-279 (1911).
P. M. S. Blackett: Proc. Roy. Soc. *102A*, pp. 1-17 (1922).
L. de Broglie: Jour. de Phys. (5) *6*, pp. 161-168 (1916).
M. de Broglie: Jour. de Phys. (6) *2*, pp. 265-267 (1921); C. R. *173*, *172* (1922).
J. Chadwick: Phil. Mag. *40*, pp. 734-746 (1920).
A. H. Compton: Phil. Mag. *46*, pp. 897-911 (1923).
J. Danysz: Le Radium *9*, pp. 1-6 (1912); *10*, pp. 4-6 (1913).
C. Davisson, C. H. Kunsman: Science, November 25, 1921 and later issues; Phys. Rev. *21*, pp. 637-649 (1923); Phys. Rev. *22*, pp. 242-258 (1923).
C. D. Ellis: Proc. Roy. Soc. *101A*, pp. 1-17 (1922).
H. Geiger and E. Marsden: Phil. Mag. *25*, pp. 604-623 (1913).
W. D. Harkins and R. D. Ryan: Jour. Am. Chem. Soc. *45*, pp. 2095-2107 (1923).
L. Meitner: ZS. f. Phys. *17*, pp. 54-66 (1923).
C. Ramsauer: Ann. d. Phys. *72*, pp. 345-352 (1923).
O. W. Richardson, R. N. Chaudhuri: Phil. Mag. *45*, pp. 337-352, and *46*, pp. 461-472, 553-564 (1923).
P. A. Ross: Phys. Rev. *22*, pp. 201-202 (1923). (Scattering of light.)
P. A. Ross: Phys. Rev. *22*, pp. 221-225 (1923). (X-ray spectra.)
E. Rutherford: Phil. Mag. *21*, pp. 669-688 (1911). (Deflection of charged particles by atom-nuclei.)
E. Rutherford and H. Robinson: Phil. Mag. *26*, pp. 717-729 (1913).
H. D. Smyth: Proc. Roy. Soc. *102A*, pp. 283-293 (1922-23); *104A*, pp. 121-131 (1923).
G. Wentzel: Phys. ZS, *23*, pp. 435-436 (1922). Ann. d. Phys. *69* (1922).
R. Whiddington: Phil. Mag. *43*, pp. 1116-1126 (1922).
C. T. R. Wilson: Proc. Roy. Soc. *104A*, pp. 192-212; 1923.

Contributors to This Issue

S. P. Shackleton, B.S., E.E., University of Michigan, 1915; American Telephone and Telegraph Company, Engineering Department, 1915–19; Department of Development and Research, 1919– . Mr. Shackleton's work has been connected with toll circuit development, including toll switchboards, telephone repeaters and testboards.

H. W. Purcell, B.S., Harvard University, 1905; Engineering Department of the Western Electric Company, 1906–20; Department of Development and Research, American Telephone and Telegraph Company, 1920– . Mr. Purcell has been associated with the design of electromagnets and more recently with the development of machine switching apparatus.

Walter A. Shewhart, A.B., University of Illinois, 1913; A.M., 1914; Ph.D., University of California, 1917; Engineering Department, Western Electric Company, 1918– . Mr. Shewhart has been engaged in the study of the relationship between the microphonic and the physico-chemical properties of carbon.

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.

Robbins P. Crowell, Southern New England Company, assistant chief operator, New Haven, 1907–09, chief operator, New London, 1909–10; New York Company, assistant manager, 1910–13; engineer in the Traffic Department, 1913–16; Engineering Department, American Telephone and Telegraph Company, 1916– . Mr. Crowell has been concerned chiefly in the preparation and introduction of peg counts and force adjustment routines.

Francis F. Lucas, Plant and Engineering Departments of Associate Companies, 1902–10, Installation Department Western Electric Company, 1910–13; Engineering Department, 1913– . Mr. Lucas is a materials engineer and has specialized in the application of microscopy to industrial problems.

J. G. Ferguson, B.S., University of California, 1915; M.S., 1916; research assistant in physics, 1915–16; Engineering Department of the Western Electric Company, 1916– .

Karl K. Darrow, S.B., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., in physics and mathematics, University of Chicago, 1917; Engineering Department, Western Electric Company, 1917– . At the Western Electric, Mr. Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

The Bell System Technical Journal

April, 1924

High Frequency Amplifiers

By H. T. FRIIS and A. G. JENSEN

IN this paper, a simplified mathematical treatment of the theory of high frequency amplifiers is presented, and the theory is verified by experiment. This method of mathematical analysis provides a

ERRATA

ISSUE OF JANUARY, 1924

On page 162, line 11 from bottom of page, and on page 163, line 1:

read $K\gamma$ *instead of* $K\delta$.

On page 173, line 3 from bottom of page:

read emerging *instead of* engineering.

at its natural frequency, i.e., the transformer inductance and distributed capacity must be in resonance. We have therefore, next treated the simplest type of resonance circuit amplifier, namely, a single tuned circuit amplifier, and it is shown that exactly the same method can be used for a choke coil amplifier or a close coupled transformer amplifier. Finally, it is shown that a loosely coupled transformer amplifier can be treated like two coupled tuned circuits.

Considering a low frequency transformer-coupled amplifier, in Fig. 1 (a) there is shown an amplifier tube I with its output transformer *T* working into another tube II and in Fig. 1 (b) is given the corresponding equivalent circuit. The equivalent circuit is obtained by the theorem, that the plate circuit of a vacuum-tube may be

treated as an ordinary a.c. circuit, consisting of the external impedance in series with a resistance R_p, and in which the impressed emf. is μe_g, R_p being the internal plate impedance of the tube, μ the amplification constant of the tube and e_g the voltage applied to the grid.

In Fig. 1 (b) C_p is the plate to filament capacity of tube I, and the input impedance of tube II is represented by a resistance R_g in parallel with a condenser C_g.

The maximum amplification which can be obtained by this amplifier is given by the well-known expression

$$K = \frac{e_o}{e_g} = \tfrac{1}{2}\mu \sqrt{\frac{R_g}{R_p}}, \tag{1}$$

but this maximum amplification can only be obtained when

$$\left.\begin{aligned} &\omega L_1 >> R_p, \\ &\omega L_2 >> R_g, \\ \text{and} \quad &\frac{\omega L_1}{R_p} = \frac{\omega L_2}{R_g}. \end{aligned}\right\} \tag{2}$$

Large reactances ωL_1 and ωL_2 can only be obtained at low frequencies because at higher frequencies the effects of internal tube capacities and the distributed capacity of the coil become large. This may best be illustrated by means of the table given below:

TABLE I

Coil No.	Inductance L	Natural Tuning Frequency f	Reactance at Half Natural Tuning Frequency $\pi f L$
1	.0025 henries	10^6 cycles	8,000 ohms
2	.25 "	10^6 "	80,000 "
3	2.5 "	10^6 "	800,000 "

The tube capacity plus the distributed capacity of each of the three coils for which these data are given is assumed to be 10 $\mu\mu f$. Since transformers in order to give a flat band must work below their natural frequency a much higher impedance than given by $\pi f L$ in the Table can therefore not be obtained. It is thus seen that only at audio frequencies is it possible to build a transformer with an impedance which is high compared with the tube resistances, the plate resistance being of the order of 6,000 50,000 ohms for ordinary receiving tubes and the grid resistance R_g being as high as 4×10^6 ohms but often limited to 500,000 ohms by an added resistance.

At higher frequencies sufficiently high impedances can only be obtained by working at the natural frequency of the transformer, and to illustrate this we shall in the following give some results of experiments made with ordinary tuned circuit amplifiers, choke coil amplifiers and loosely coupled transformer amplifiers at high frequencies.

Tuned Circuit and Choke Coil Amplifiers

In Fig. 2 there are shown to the left two different ways of connecting up a tuned circuit amplifier, and to the right are given the corresponding equivalent circuits. The input impedance to the next tube is assumed to be a pure resistance R_g, thus neglecting the grid-

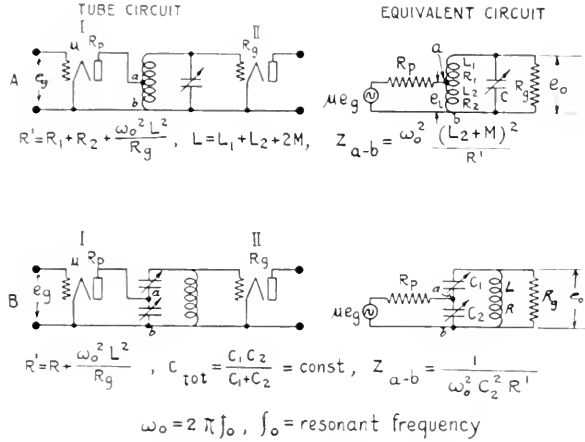


Fig. 2—Schematic of Tuned Amplifier Circuits

filament capacity and the grid-plate capacity of this tube. The effect of the grid-filament capacity, however, will only be to detune the circuit a little and can, therefore, be compensated for by retuning the condenser C (or C_1 and C_2) and the effects of the coupling through the grid-plate capacity of the second tube will be treated specially later.

Fig. 2 gives the well-known formulas for the equivalent series resistance R' of the circuit at resonance and for the impedance of the circuit Z_{a-b} measured between points a and b at resonance.

From this we then get in Case (A)

$$e_i = \mu e_g \frac{Z_{a-b}}{Z_{a-b}+R_p} = \mu e_g \frac{\omega_o^2(L_2+M)^2}{\omega_o^2(L_2+M)^2+R_pR'} \tag{3}$$

and, assuming that $R_p >> \omega_o(L_2+M)$,

$$e_o = e_i \frac{L}{L_2+M}.$$

Hence, defining the voltage amplification K of the first stage as the voltage impressed upon the grid of tube II divided by the voltage impressed upon the grid of tube I, we have

$$K = \frac{e_o}{e_g} = \mu \frac{\omega_o^2 L(L_2+M)}{\omega_o^2(L_2+M)^2+R_pR'}. \tag{4}$$

In order to find the step-up ratio, which gives maximum amplification we have to solve for L_2+M in the equation $\delta K \ \delta(L_2+M)=0$, which gives

$$R_p = \frac{\omega_o^2(L_2+M)^2}{R'} = Z_{a\ b}, \tag{5}$$

and by inserting this in equation (4) we get

$$K_{max} = \frac{\mu}{2}\left(1 - \frac{\omega_o L}{R'}\right). \tag{6}$$

From equation (5) it is seen that the condition for maximum voltage amplification is exactly the same as the well-known condition for maximum power amplification; namely, that the external impedance $Z_{a\ b}$ inserted in the plate circuit must be equal to the internal tube impedance R_p.

By repeating the calculations given above for Case (B) in Fig. 2, it will be found that this condition again holds good, and also it will be found that the expression for K_{max} is the same.

As already mentioned the resistance R' in the formulas above includes the equivalent series resistance introduced in the tuned circuit by the impedance of the input circuit of tube II, but in many cases this extra resistance will be negligible as compared to the resistance of the coil itself, and equation (6) thus gives us the very interesting

information that the maximum amplification obtainable with a tuned circuit amplifier is proportional to the *ratio of the inductive reactance to the square root of the resistance*. In the case of an ordinary selective circuit such as a tuned loop antenna the output voltage developed is proportional to the ratio of the inductive reactance to the *first power of the resistance*. This does not mean that low resistance is less desirable in amplifier coils than in ordinary tuned circuits but it does mean that the penalty exacted by increasing the resistance is not as great.

In order to test the formulas given by equations (5) and (6), a series of experiments have been carried out.

For measurements of the maximum amplification of a tuned circuit amplifier a circuit as shown in Fig. 3 was used.[1] The grid of the am-

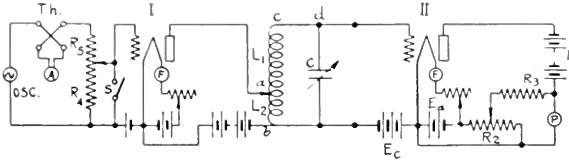


Fig. 3—Method of Measurement of Tuned Amplifier

plifier tube I is connected up to a known resistance, through which is passed a known current, and the voltage across the tuned circuit is measured by means of the tube-voltmeter II. The inductance L_1L_2 is made up of a single layer solenoid closely wound with 173 turns of solid wire and its value was 1.63×10^{-3} henries.

Keeping the frequency and the input from the oscillator constant the circuit is tuned to resonance by means of the variable condenser C and the lead from the plate to the coil is then moved along the coil until a point is reached which gives maximum deflection of the tube-voltmeter. During this process it is necessary to retune the circuit for each new point tried. Having thus obtained the right step-up for a certain frequency we then measure the amplification for different frequencies and get the amplification curves shown in the upper half of Fig. 4. On the lower half of Fig. 4 are given the number of turns (L_2) across the plate of the amplifier tube and also the capacity of the condenser C for each of the four cases shown.

[1] For a more detailed description of the method of measurement, see section titled "Measurements" below.

In order to calculate the maximum amplification from formula (6) it is necessary to know the resistance R' of the circuit, the voltage amplification factor μ, and the internal plate impedance R_p of the amplifier tube. R' was obtained by running resonance curves for the circuit with the tubes connected up as usual, but with no filament current in the amplifier tube, in which case R_p may be regarded as being infinite.

These resonance curves are shown in the lower part of Fig. 4, and the resistance is then calculated from the well-known formula

$$R' = 2\pi(f_1 - f_2)L, \tag{7}$$

in which f_1 and f_2 are the frequencies, for which $E = E_{max} \sqrt{2}$.

The resistance R' may also be obtained from the amplification curves as these can be regarded as resonance curves for the tuned circuit with the resistance R_p across part of the coil, and since this

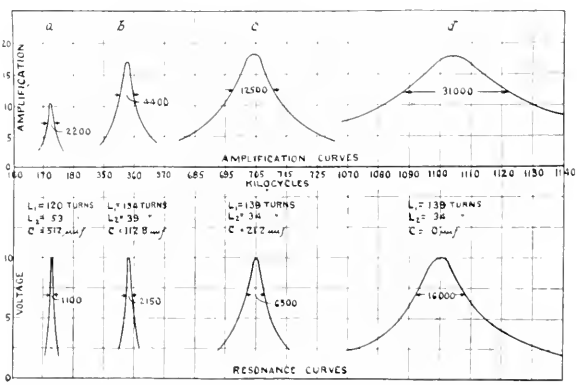


Fig. 4 Experimental Amplification and Resonance Curves of Tuned Circuit Amplifier

part of the coil is chosen so as to give $Z_{a\ b} = R_p$, the equivalent series resistance should have increased to exactly twice the value found in formula (7). By comparing the widths of the amplification curves in Fig. 4 with the widths of the corresponding resonance curves it is seen that this actually was the case.

The internal plate impedance R_p and the amplification factor μ were obtained from the slope of the static characteristic of the amplifier tube used (a Western Electric 215-A or "peanut" tube).

The results of the calculations are given in Table II and the calculated values of K_{max} are seen to agree very well with the measured values given in the last column.

TABLE II

$L = 1.63 \times 10^{-3}$ henries, $R_p = 22{,}000$ ohms, $\mu = 6.4$

Frequency	$f_1 - f_2$	$R' = 2\pi L(f_1 - f_2)$	Calculated Maximum Amplification $K_{max} = \frac{\mu}{2} \frac{\omega L}{\sqrt{R_p \cdot R'}}$	Measured Maximum Amplification
472,000	1,110	11.4	10.7	10.3
857,200	2,150	22.1	15.9	16.6
704,000	6,500	66.6	18.1	18.1
1,100,000	16,000	164	18.1	17.8

The amplification of the amplifier was also measured with no step-up, i.e., with the plate of the amplifier tube connected across the whole

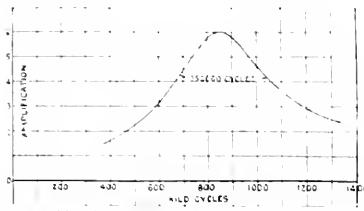


Fig. 5—Amplification Curve of Choke Coil Coupled Amplifier

coil and the tuning condenser C omitted (choke coil amplifier) the amplification curve shown in Fig. 5 being obtained.

From a resonance curve the following value is obtained:

$$R' = 2\pi L(f_1 - f_2) = 2\pi \times 1.63 \times 10^{-3} \times 15{,}000 = 153 \text{ ohms},$$

and, therefore,

$$R_{tot} = R' + \frac{\omega^2 L^2}{R_p} = 153 + \frac{4\pi^2 853{,}000^2 \times 1.63^2 \times 10^{-6}}{22{,}000}$$

$$= 153 + 3320 = 3523 \text{ ohms},$$

which inserted in formula (7) gives

$$f_1 - f_2 = \frac{R_{tot}}{2\pi L} = \frac{3523}{2\pi \times 1.63 \times 10^{-3}} = 345{,}000 \text{ cycles}$$

while the amplification curve gives $f_1 - f_2 = 350{,}000$ cycles.

As a final check of formula (6) by means of this tuned circuit, the maximum amplification was measured at 170,000 cycles with different values of extra resistance, R_{ext}, inserted in the circuit between c and d in Fig. 3. The results of these measurements agree very well with the formula as will be seen from Table III. For $R_{ext} = 160$ it was found necessary to connect the plate across the entire coil in order to get maximum amplification and thus a further increase of R_{ext} beyond 160 ohms will make it impossible to obtain maximum amplification with this circuit.

TABLE III

$f = 170{,}000$ cycles, $L = 1.63 \times 10^{-3}$ henries, $R_p = 22{,}000$ ohms, $\mu = 6.1$.

R_{ext}	Total R'	Calculated Maximum Amplification $K_{max} = \frac{\mu}{2} \sqrt{\frac{\omega L}{R_p \cdot R'}}$	Measured Maximum Amplification
0	11.6	10.2	10.7
10	21.6	7.8	7.7
20	31.6	6.4	6.4
40	51.6	5.2	5
80	91.6	3.95	3.7
160	171.6	2.85	2.7

The variation with frequency of the resistance of the coil is shown on Fig. 6. These resistance values are obtained from the resonance curves in Fig. 4, and hence indicate also the losses in the variable condenser and the loss due to the input impedance R_g.

The curve in Fig. 6 gives what may be called the "true" resistance of the circuit, which is to be distinguished from the "apparent" resistance of the circuit as measured for instance by the well-known resistance variation method. By this latter method, the resistance of the coil is assumed to be equal to such an amount of extra resistance, as inserted in the circuit will decrease the resonance current to half its former value, but this assumption is only true when the distributed capacity of the coil is negligible as compared with the capacity of the variable condenser C, or when the resistance is introduced in the center of the coil.

It follows from formula (6) that for a given coil the maximum amplification is proportional to $\sqrt{\frac{\omega L}{R'}}$, and the measurements mentioned above seem to indicate that the maximum of this ratio has already been passed in the last case (d, Fig. 4) when the coil is used simply

as a choke coil or auto-transformer (without any extra condenser). This, however, will depend upon the kind of wire used in making the coil. The coil used in the measurements above was made of No. 28

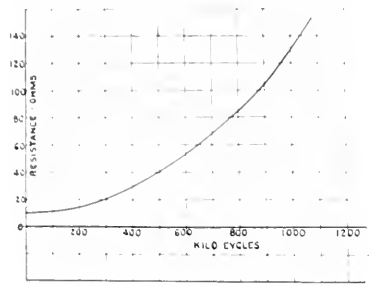


Fig. 6 Effective Resistance of Choke Coil

solid wire, but earlier results obtained by other investigators have shown that solid wire is superior to stranded wire at high frequencies, and thus it may be expected that the maximum of the ratio $\frac{\omega L}{R'}$ for a given inductance will occur at a lower frequency when the coil is made of stranded wire.

For constant frequency the maximum amplification is proportioned to the ratio $\frac{L}{R'}$ as already mentioned. It is thus desirable to adopt a construction for the coil, which will increase L without increasing $\sqrt{R'}$ proportionally. The highest amplification will in general be obtained when L is as large as possible for the frequency in question; in other words, it will be possible to obtain a higher amplification when the tuning condenser in the tuned circuit amplifier is reduced to zero, giving a simple choke coil amplifier.

For a tuned circuit amplifier with an ordinary good inductance coil made of stranded wire and of an inductance of, for instance, 200 microhenries and a high-frequency resistance of about 5 ohms, the amplification at 800 kilocycles will not be higher than about 9 times, according to formula (6) (using the same kind of tubes as in the experiments above), while with a choke coil an amplification as much as 18 times was obtained. This means that in order to get high amplification, small coils made of fine, solid wire and with large inductance and small distributed capacity should be used, rather than large

coils made of stranded wire and with smaller inductance, but with larger distributed capacity.

In practice, it is not important to go to extremes in order to reduce the distributed capacity by one or two $\mu\mu f$. because the coil will always be shunted by the tube capacities, which are of the order of 10 $\mu\mu f$. It may be mentioned that the distributed capacity of the coil used in the above experiment is 3.5 $\mu\mu f$. This means that the constructional details of such a coil are not very important, and the coil may be made as a single layer coil or as a coil wound in one or several sections of rectangular or square cross-sections, but in all cases it will be found that coils of the same inductance will have very closely the same resonance frequency provided that the same tubes and leads are used in all cases.

Some experiments made with a choke coil (or auto transformer) at about 50,000 cycles show that the formulas given above may be also used here.

The coil used in these experiments was wound on a core of iron dust and made with square cross-section. The total inductance of the coil was .33 henries and provisions were made so that the plate of the amplifier tube could be tapped across any part of the coil.

The circuit diagram was the same as that given in Fig. 3 with the exception that the condenser C was omitted. The maximum amplification curve for this coil, used as a choke coil, is given by Fig. 7, curve A. The step-up ratio necessary to obtain maximum amplification was 1:16; i.e., the plate was connected across 1 16 of the total number of turns.

The resistance of the coil is obtained from a resonance curve as before:

$$R' = 2\pi L(f_1 - f_2) = 2\pi \times .33 \times 1300 = 2700 \text{ ohms},$$

and inserting this in formula (6) gives:

$$K_{max} = \frac{6.1}{2}\sqrt{\frac{2\pi \times 51800 \times .33}{22{,}000 \times 2700}} = 15$$

while the experiment gave 14.5.

On Fig. 7 are also given the amplification curves B, C and D for a step-up of 1:4, 1:1 and 1:18, respectively.

In the two cases B and C, the selectivity of the circuit is determined almost entirely by R_p, the resistance of the circuit itself being negligible, while in case D the selectivity is practically determined by the resistance of the coil itself.

It is seen that the amplification curve C for a step-up ratio of 1:1 is extremely flat as compared with the amplification curve shown in Fig. 5 for a choke coil working at 850 kilocycles.

In connection with these experiments with tuned circuits and choke coils it may be mentioned that in order to separate the *DC* plate voltage from the *DC* grid voltage, it will often be found of ad-

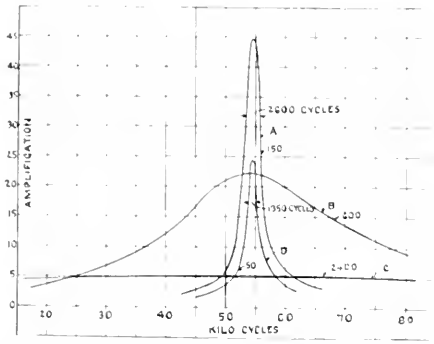


Fig. 7.—Curve Showing the Effect of Ratio of Transformation in the Characteristic of a Choke Coil Coupled Amplifier

vantage to replace the coil by a transformer with very close coupling. In all our experiments, we have found that the amplification curves obtained in the two cases are identical when the coupling coefficient for the two windings of the transformer is nearly unity.

Loosely Coupled Transformer Amplifier

From the amplification curves obtained with choke coils, it will be seen that the frequency range obtainable with a choke coil amplifier is not as wide as might be desirable in some cases. This is especially true for higher frequencies between 300,000 and 1,000,000 cycles, and where a wide frequency band is desired these choke coils have, therefore, been replaced by transformers with a rather loose coupling, in which case the transformers will have the characteristics of two ordinary coupled circuits and give an amplification curve with two peaks.

It has been found by experiment that such transformers can actually be treated just as ordinary coupled circuits and the amplification

curves can be computed by means of the well-known formulas for current and voltage conditions in two coupled circuits.

Before going into the details of these experiments, it is worth while to consider briefly the general relations involved as indicated by the curves obtained with two coupled circuits, each tuned to 52,000 cycles. These curves are shown in Fig. 8. The coils used

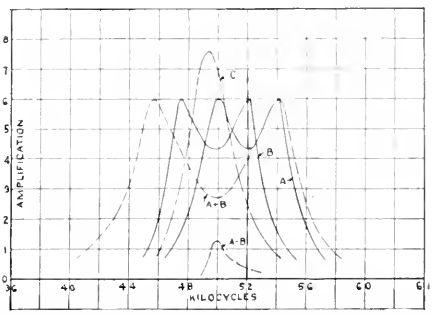


Fig. 8 – Curves Showing the Effect of Coupling Inductive and Capacitive, on Amplification Characteristic of Coupled Tuned Circuits

had an inductance of 10 millihenries and were tuned by condensers. The circuit of the apparatus employed in obtaining the curves is given in Fig. 9.

Curve A gives the amplification for inductive coupling alone.

Curve B is for capacitive coupling alone.

Curve $A+B$ is for both capacitive and inductive coupling aiding each other, each coupling having the same value respectively as in curves A and B.

Curve $A-B$ is with the two couplings opposing each other and

Curve C is the same as $A-B$ but with different value of the inductive coupling.

The curves have the same shape as the well-known resonance curves for two coupled circuits with the oscillator input in series with the primary circuit, where the peak frequencies are given by the following approximate formulas:

$$\text{Inductive coupling: } f' = \frac{f_o}{\sqrt{1-k}},\quad f'' = \frac{f_o}{\sqrt{1+k}},$$

Capacitive coupling: $f' = f_1 \sqrt{\frac{C}{C+2C'}}$

where $f_1 = \frac{1}{\sqrt{LC}}$, $k = \frac{M}{L}$ = coefficient of coupling.

Having thus demonstrated the general shape of the amplification curves for a two coupled tuned circuit amplifier, the action of a loosely coupled transformer amplifier for high frequencies will be treated.

The transformer used in this experiment was made up of two similar pancake coils, 2" diameter, wound with 210 turns of solid wire. Fig. 9 shows the circuit diagram. Curves *A* and *B* in Fig. 10a

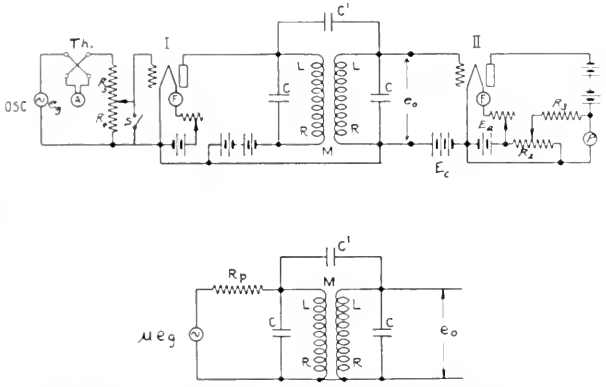


Fig. 9—Method of Measurement of Loosely Coupled Transformer Amplifier

show the measured amplification curves for a 3 8" distance between the windings. The coupling condensers *C'* were omitted but even then there was some capacity coupling left due to the distributed capacity between the coils. The curves *A* and *B* correspond respectively, to an aiding and an opposing action of this capacitive coupling. Interchanging the leads to either coil changes the amplification curve from one type to the other.

The self inductance and mutual inductance of the coils were measured at low frequency and found to be:

$$L = 2.1 \times 10^{-3} \text{ henries}, \quad M = .95 \times 10^{-3} \text{ henries}.$$

The resistance of the coils was measured as described before by taking resonance curves at different frequencies. The distributed capacity of each coil was 11×10^{-12} farad (including tube capacity). By means of these values, the curve C was calculated.[2] The unknown capacity coupling makes it impossible to predict the exact shape of a transformer coupled amplifier from the constants of the circuits. However, the calculated curve C (calculated for inductive coupling only) will give a general idea of the shape of an experimental curve A.

Curves A and B, Fig. 10b, show the amplification curves for the case of capacity coupling alone. A is the experimental and B the calculated curve and they are seen to give fair agreement. The coupling capacity was 21×10^{-12} farad and the distributed capacity of the coils was 19.3×10^{-12} farad, the increase, as compared with the case of inductive coupling, being due to the ground capacities of the coupling condenser.

In connection with this type of amplifier it may be mentioned that a higher amplification naturally can be obtained if the plate of the amplifier tube is connected across a part of the primary circuit only, maximum amplification corresponding to the circuit impedance being equal to the plate impedance. However, the same effect will take place here as was shown for the tuned circuit amplifier, namely, that the band width will decrease with increase in amplification. Using transformers at their natural frequency instead of coupled tuned circuits with outside condensers will give broader bands or higher amplifications corresponding to the single tuned circuit amplifier.

[2] The following two formulas have been used for calculating the amplification of the circuit shown in Fig. 9.

Inductive Coupling:

$$\text{Amplification} = \frac{e_s}{e_g} = \frac{\omega M}{Z}\,\frac{1}{\omega C R_p(1-\omega^2 L_1 C + R_1 + j\omega L_1 + R_p R_1 \omega C)}\cdot\mu$$

where $Z = \sqrt{R_2^2 + \left(\omega L_2 - \frac{1}{\omega C}\right)^2}$, $R_1 = R\left(1 + \left(\frac{\omega M}{Z}\right)^2\right)$ $\omega L_1 = \omega L - \left(\frac{\omega M}{Z}\right)^2\left(\omega L_2 - \frac{1}{\omega C}\right)$

Capacitive Coupling:

$$\text{Amplification} = \mu\,\frac{R + j\omega L}{A + iB}$$

where $A = R_p\left(2 - 2\omega^2 L C + 2\frac{C}{C_g} - \omega^2 L C\frac{C}{C_g}\right)\frac{1}{R^2 + \omega L^2}\frac{\omega L}{\omega C_g} + R\left(1 + \frac{C}{C_g}\right)$,

$$B = RR_p\omega C\left(2 + \frac{C}{C_g}\right) - \frac{RR_p}{\omega C_g(R^2 + \omega^2 L^2)} + \omega L\left(1 + \frac{C}{C_g}\right) - \frac{1}{\omega C_g}.$$

Amplifiers with Several Stages. "Feed Back" Action

The experiments so far have shown, that with one stage of amplification and with the amplifier working into a detector tube without grid condenser and leak, it is always possible to calculate the amplification curve from the constants of the tubes and of the coils, regardless

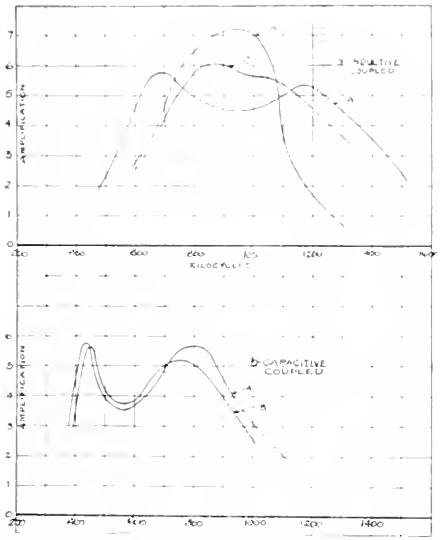


Fig. 10. Amplification Curves of a Loosely Coupled Transformer Amplifier Showing Effect of Coupling

of whether the connection between the amplifier and the detector consists of a simple tuned circuit, a choke coil, two coupled circuits or a loosely coupled transformer, the circuits being treated simply as ordinary tuned circuits.

Also the experiments have shown that a higher amplification can be obtained in the 50,000 cycles region than around 1,000,000 cycles, as might be expected from formula 6.

The next question is: What happens when more than one stage of amplification is used? If, for instance, the amplification for one stage is 10, will then the amplification for two stages be 100 or, in other

words, in a multiple-stage amplifier is it possible to get the total amplification curve from the curve for the amplification per stage by multiplying them together?

The answer to this question is that the total amplification of a multi-stage amplifier will, in general, be lower than the value obtained by multiplying the amplification values per stage, and the reason for this is to be found in the input impedance of the tubes. So far, we have assumed the input impedance of the tube after the amplifier to be high as compared with the impedance of the tuned circuit (or transformer) and this is correct for a plate curvature detector, in which the impedance of the load in the plate circuit is negligible at the frequency of the amplified current but if the next tube is another amplifier it is only true at lower frequencies. It has been shown[3] that the input impedance of a vacuum tube can readily be calculated by means of the constants of the tube and the output impedance.

For the tubes used in the foregoing experiments we have the following approximate constants:

C_{g-p} = Grid to plate capacity $= 3 \times 10^{-12}$ farad
C_g = Grid to filament capacity $= 5 \times 10^{-12}$ farad
C_p = Plate filament capacity $= 5 \times 10^{-12}$ farad
R_p = Plate impedance $= 20{,}000$ ohms
μ = Amplification constant $= 6$.

The output impedance including the plate-filament capacity will be assumed to be a resistance equal to the plate impedance.

If the input impedance is represented by an apparent resistance R'_g in parallel with an apparent capacity C''_g, we get for R'_g and C''_g the values given in Table IV.

TABLE IV

$$R'_g = \frac{\omega^2 R_p^2 C_{g-p}^2 + 1}{\omega^2 R_p C_{g-p}^2 (\mu + 2)} \qquad C''_g = C_g + 2C_{g-p} - \frac{\mu + 2}{\omega^2 R_p^2 C_{g-p}^2 + 1}$$

Frequency	R'_g	C''_g
10^6 cycles	$7 \cdot 10^6$ ohms	17×10^{-12} farad
10^5 "	7×10^5 "	17×10^{-12} "
10^5 "	$7 \cdot 10^5$ "	17×10^{-12} "
10^5 "	$7 \cdot 5 \cdot 10^3$ "	16×10^{-12} "

[3] H. W. Nichols, *Phys. Rev.*, Vol. 13, p. 405, 1919. John M. Miller, Bureau of Standards "Scien. Pap. No. 351, 1919.

From this table it is seen that the effect of the input impedance is negligible at frequencies up to about 100,000 cycles, but for frequencies in the broadcasting range, the input impedance will introduce an appreciable loss in the preceding circuit, which will result in a drop in amplification below the value obtained for a single stage amplifier. It is seen that the input impedance R'_g for broadcasting frequencies

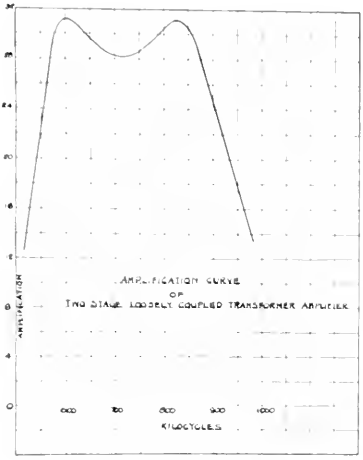


Fig. 11. Amplification Curve of Two Stage Loosely Coupled Transformer Amplifier

is of the same order of magnitude as the plate impedance R_p, which means that it will be of no advantage to use much step-up in choke coils or tuned circuits for an amplifier with more than one stage, since the amplification in no case will be much higher than u per stage, except for the last stage, which is working into the detector.

The loosely coupled transformers of the type already discussed will, on the other hand, work very well in a two-stage amplifier, since there is no step-up used in these, and the amplification will be very nearly twice the amplification for a one-stage amplifier, as will be seen from the amplification curve shown in Fig. 11. The width of such an amplification curve can be increased by proper adjustment of the transformer inductances but the amplification will naturally drop correspondingly.

The values of R'_g and C'_g given in Table IV were calculated on the assumption of a pure resistance load R in the plate circuit. If the load in the plate circuit is an impedance $Z = R + j\,x$, it will be found that the sign of the apparent shunt resistance R'_g will depend upon the

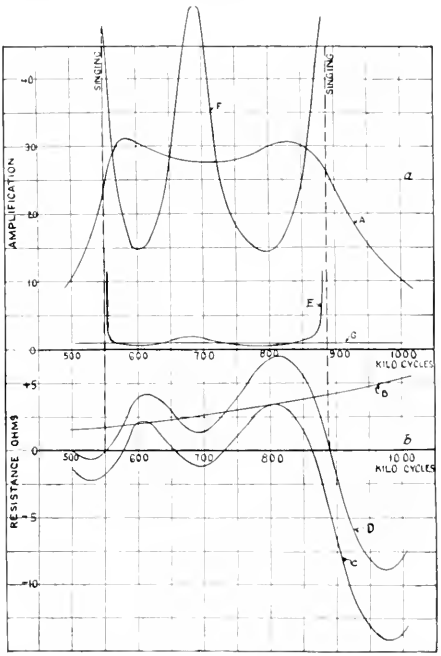


Fig. 12—Total Amplification of Transformer Coupled Receiver and Effect of "Feed Back" on Loop Resistance

sign of the reactance x. For a capacitive load, the resistance R'_g will always be positive, but for an inductive load, R'_g may in some cases become negative and we then have "feed back" or regeneration occurring through the tube. The negative resistance introduced in the circuit below the resonance frequency may in certain cases be so high that it more than neutralizes the positive resistance of this circuit which means that the set will start to oscillate or "sing."

As an illustration of the effect of this "feed back" action, there are given in Fig. 12 some curves obtained for a two stage high frequency amplifier with loosely coupled transformer stages. The input circuit to the amplifier consisted of a loop antenna circuit tuned to the frequency of the induced signal.

Curve *A* shows the straight high frequency voltage amplification of the set, as measured with resistance input to the grid of the first high frequency amplifier. (Same as curve shown in Fig. 11.)

Curve *B* gives the actual resistance of the loop used with the set.

Curve *C* gives the resistance introduced in the loop due to "feed back" action from the first stage.

Curve *D* gives the resulting apparent resistance of the loop (Curve *B*+Curve *C*) and

Curve *E* shows the "feed back" amplification of the set. (Curve *B:* Curve *D*.)

Curve *F* shows the total amount of amplification obtained by the set which is the product of the ordinary voltage amplification (Curve

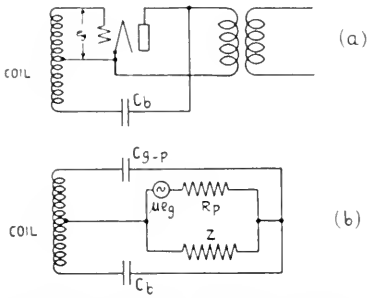


Fig. 13 Schematic of Balancing Condenser Action

A) and the "feed back" amplification (Curve *E*) and it is thus seen that the feed back action makes the total amplification vary irregularly in a very undesirable manner, and also makes the set "sing" at certain frequencies.

In order to avoid this, it is necessary to provide some means of balancing out the effect of the grid plate capacity of the tubes, and Fig. 13 (a) shows how this may be done[4]. The filament of the tube

[4] See Patent No. 1,183,875 issued to R. V. L. Hartley, and Patent No. 1,334,418 issued to C. W. Rice.

is connected to the middle of the coil, the grid to one end and the plate is connected through a small balancing condenser to the other end of the coil. In Fig. 13 (b) is given a schematic diagram of the circuit, which shows that the effect of C_b upon the coil circuit is just opposite the effect of C_{g-p}, so that the circuit can be regarded as an ordinary bridge circuit. It will, therefore, always be possible by proper adjustment of the condenser C_b to neutralize the effect of the feed-back action as shown by curve G in Fig. 12.

The same kind of an arrangement can be used between the different stages in a multi-stage high frequency amplifier, and it is thus seen that by proper use of such balancing condensers, it will be possible to obtain for a multi-stage amplifier a total amplification which is practically equal to the product of the amplifications per stage. This is true for a multi-stage tuned circuit coupled amplifier but for transformer coupled amplifiers, where it is more difficult to obtain a 180° phase difference of voltages, the advantage of the balancing condenser is not so great.

Of course, this favorable result presupposes that the wiring of the amplifier is properly done and the different stages shielded carefully from each other so that no external coupling exists between them.

Resumé

What has been said about amplifiers in the preceding sections can be summarized as follows:

With a given type of amplifier the same general shape of the amplification curve is obtained regardless of the frequency range at which the amplifier is designed to operate.

Thus, a low amplification over a wide frequency range will be obtained by using loosely coupled transformers or choke coils without any step-up, while a high amplification over a narrow range of frequencies can be obtained by using choke coils or tuned circuits with a proper step-up. In this last case, it will be necessary to use a small tuning condenser across the coils in order to make the frequency range of the amplifier wider, and the higher amount of amplification is, therefore, obtained only by a sacrifice of tuning facilities of the set. In a multi-stage amplifier it may, however, often be found of advantage to use a combination of low amplification stages and high amplification stages so that, for instance, one tuned circuit stage with high step-up and variable condenser is used in connection with one or several stages of choke coils or loosely coupled transformers with low amplification and a wide frequency range. The maximum amplifi-

cation obtained with any kind of an amplifier will, in general, be higher at the lower frequencies, due to the lower loss and the higher ratio of L over C obtainable.

The width of the frequency band for a choke coil amplifier will be smaller, the higher the frequency due to the decrease in ωL with increasing frequency, and at broadcasting frequencies it will, therefore, in general be found advantageous to use loosely coupled transformers rather than choke coils, whenever a wide frequency band is desired. In addition to giving a wider frequency band, lower frequency amplifiers have the advantage of a smaller grid-plate feed back action.

Amplification Measurements at High Frequencies

In order to make a thorough study of radio frequency amplification, it is necessary to have a dependable method of measurement. Such a method developed in our laboratory and used very successfully will be described here.

In order to obtain an accurate comparison between different types of amplifiers, in which any type of resonant coupling is used, it is essential that these amplifiers be operated from a resistance input and not from an input containing a tuned circuit. With a tuned circuit it is not only very difficult to obtain an accurate measure of the voltage impressed upon the amplifier but considerable regeneration may occur between this input circuit and the output circuit of the first amplifier tube. There is, naturally, also a feed-back action in connection with a resistance input circuit, but its effect is negligible when the resistance is only a few hundred ohms. When the characteristic of a radio frequency amplifier with a resistance input has been accurately determined, its characteristic when used with a tuned circuit input may be determined as will be described later.

A schematic circuit diagram of the apparatus as used is shown in Fig. 3. To the left is shown the input apparatus which consists of an oscillator, a sensitive thermocouple and a potentiometer. The drop across the resistance R_4 of the potentiometer is used as the input to the amplifier stage I. The output of the amplifier stage is measured by the tube voltmeter II shown to the right in Fig. 3. The tube voltmeter II may be a low frequency detector in the case of amplification measurements of an actual receiver set.

It is necessary first to calibrate the tube voltmeter or detector II which is done by disconnecting it from the amplifier and connecting it directly across the potentiometer $R_4 - R_5$. R_4 is then adjusted to,

say, 500 ohms and the current through it adjusted to some convenient value, such as 1 milliampere. This voltage of .5 volt will be sufficient with most tubes to give a change in the plate current of 30 to 40 microamperes.

The tube voltmeter is then reconnected to its normal place in the circuit and the resistance R_4 is connected to the input of the amplifier. Keeping the current constant at the value of 1 milliampere, the resistance R_4 is adjusted until the change in the detector plate current is the same as before. It is immediately apparent that the amplification will be the ratio of the known voltage on the grid of the detector, that is .5 volt, to the voltage on the input of the amplifier, as indicated by the product of the resistance R_4 and the current through it. The current having been kept constant, the amplification is the quotient of the 500 ohms used when calibrating the detector and the resistance value obtained with the amplifier included.

Considerable precaution must be observed to make sure that no energy is getting into the amplifier circuit except that which may be measured by the voltage drop across the resistance R_4. This necessitates the most careful shielding especially when the amplification is more than 50 times.

With the measuring apparatus described a dependable input voltage as small as 1 millivolt can be obtained. The maximum amplification which can be measured directly is, therefore, of the order of 500 times when the output voltage to the detector is of the order of one half of a volt.

For the measurement of higher amplification the following indirect method may be used.

The amplification is artificially decreased in some manner such as reducing the number of stages in the circuit and this reduced amplification is measured in the usual manner. The input current is then reduced and the input resistance increased keeping the plate current of the detector constant, the voltage impressed on its grid being determined by the previous calibration. The amplification is now increased to its normal value and the input resistance decreased until the detector plate current has its original value. The ratio of decrease in input resistance will thus give the increase in amplification and the total amplification will be the product of this and the smaller amplification as first measured.

The smaller current through the input resistance, which is obtained by this method and which will generally be less than can be determined by the most sensitive thermocouple, will reduce the pick-up to a sufficiently low value to give satisfactory results. In this connec-

tion it may be noted that an excellent test for the presence of undesirable pick-up is the closing of a switch (S) placed at the input of the amplifier. With this switch closed there should be no appreciable input to the detector.

Direct high frequency amplification measurements require input units made up of very carefully constructed attenuation boxes or potentiometers and well shielded oscillators. Such units have been developed in connection with field measurements and are described in a paper on "Radio Transmission Measurements," by Messrs. Bown, Englund and Friis and "Note on the Measurements of Radio Signals," [5] by Englund.

On the right in Fig. 3 is shown, as mentioned before, the circuit diagram of a "tube voltmeter" such as is used in many high frequency measurements. The tube voltmeter is essentially a plate current curvature detector. The grid is made negative by means of the grid battery E_c, so that the normal plate current of the tube is very small (of the order of 50 microamperes or so), and this plate current is further balanced out by means of the potentiometer arrangement R_2, R_3, so that the plate current meter reads zero when the input to the tube voltmeter is short-circuited. This arrangement has the advantage of making it possible to utilize the entire scale of the meter and to obtain the measured voltage from a single reading instead of the difference of two readings. Such a tube voltmeter built with an "X" tube will give a deflection of 1 microampere for an input voltage of about 1 5 of a volt, and the calibration will stay remarkably constant for several months and is *independent* of the *frequency* at which it is calibrated. The values of the resistances in the resistance boxes used at high frequencies may, therefore, be checked by using the boxes for calibrating a tube voltmeter first at 60 cycles and afterwards at, for instance, 1,200 kilocycles. If the two calibration curves obtained are exactly identical, then the resistance has not changed appreciably within this frequency range.

In measuring the amount of "feed-back" amplification in a receiving set, it is not possible to use a method as direct as described above. The "feed-back" or regeneration in a set is, as already mentioned, due to the coupling between the grid circuit and the plate circuit of the tubes through the grid-plate capacity, and will depend upon both the load in the plate circuit and the nature of the input circuits. If, for instance, it is desirable to measure the amount of "feed-back" amplification due to the coupling between the loop circuit and the

[5] Proc. Inst. R. E., Vol. 11, No. 1, February, 1923. Proc. Inst. R. E., Vol 11, No. 2, April, 1923.

plate circuit of the first amplifier in a high frequency amplifier set, it will not be possible to measure this with a resistance input to the amplifier since in this case the "feed-back" has no appreciable effect. In order to get the correct value for the "feed-back" amplification, the set must be connected up to the same loop with which it is going to be used and the measurements can then be made in the following way.

A resistance box is inserted in the middle of the loop and a tube voltmeter is connected across half of the loop in addition to the receiving set as shown in Fig. 14. With the filament circuit of the set open, a strong high frequency emf. is induced in the loop and the loop circuit is tuned until the tube voltmeter reads a maximum.

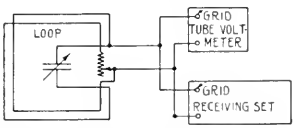


Fig. 14—Method of Measuring "Feed Back" Amplification

A "feed-back" action in the set will then produce a change in the tube voltmeter reading when the filament current is switched on. If the "feed-back" action is positive, i.e., if the resistance introduced in the loop is negative, then the tube voltmeter reading will increase, and in order to bring it back to its former value, the resistance of the loop is increased by an amount R' by means of the resistance box. If, on the other hand, the "feed-back" action is negative, the resistance of the loop must be decreased in order to obtain the former value of the tube voltmeter reading.

R' represents the equivalent series resistance introduced in the loop circuit by the "feed-back" action, and the apparent resistance of the loop is, therefore, $R-R'$, where R is the actual resistance of the loop. The voltage impressed upon the grid of the first tube is inversely proportional to the apparent resistance of the loop, and the amount of "feed-back" amplification is, therefore, defined as the ratio $K'=R/(R-R')$ where R' must be taken with the proper sign. This ratio is seen to be a direct measure of the increase (or decrease) in input voltage due to the "feed-back" action in the set and the total amount of amplification in a set at a certain frequency will then be given by the product of the ordinary voltage amplification factor K and the "feed-back" amplification factor K'.

In determining K', it is necessary to know the actual resistance R of the loop and this may be conveniently obtained by the reactance variation method using a tube voltmeter across half of the loop as the voltage indicating device. It has been found that the loss introduced by such a tube voltmeter is negligible, a fact which can be easily checked by connecting two similar tube voltmeters across the loop and determining the maximum reading of one of them. When the other one is then disconnected and the loop condenser slightly readjusted so as to again give maximum reading of the first tube voltmeter, it will be found, that the two readings obtained are exactly the same.

The discussion of the two types of amplification measurements of high frequency amplifiers may be summarized as follows:

The *ordinary voltage amplification* K is defined as the ratio of the amplified signal voltage impressed on the grid of the low frequency detector and the signal voltage impressed on the grid of the first amplifier tube. This amplification is measured by using a resistance input to the amplifier and includes the effect of "feed-back" action between the stages in the amplifier. This "feed-back" action between stages can naturally be analyzed by a method similar to the one used to determine the "feed-back" action between the amplifier and its tuned input circuit.

The *"feed-back" amplification* K' is defined as the increase (or decrease) of signal voltage due the "feed-back" action between amplifier and its tuned input circuit. The "feed-back" amplification depends upon the selectivity of the input circuit and will only vary slightly from unity when the resistance of this circuit is very large, while large variations, as shown in Fig. 12, may be found when a selective input circuit is used.

The total amplification is defined as the product of the ordinary amplification K and the "feed-back" amplification K'.

Design Characteristics of Electromagnets for Telephone Relays

By D. D. MILLER

NOTE: The electromagnets described are confined to relays, although the principles involved apply as well to selector magnets, clutch magnets and electromagnets in general. A treatment from the viewpoint of the telephone engineer is given of the important considerations which determine the design of the magnetic parts of relays and the economics of the winding dimensions. A knowledge of these factors as well as of the general considerations which are discussed is of great importance in the selection and application of relays to the telephone system. The operating and economic importance to the Bell System of the great number of relays required in the operation of the plant has been described in a previous paper.[1]

INTRODUCTION

ELECTROMAGNETS or relays as generally used in telephone switchboards are simply switches which are controlled electromagnetically. These switches may be required to open or close a number of separate and distinct circuits simultaneously or in a certain sequence. In many cases it is essential that the relay switch be opened or closed very quickly as this time may have a direct influence on the amount of apparatus required and consequently the first cost of the plant. The operating time of the relays also has a direct influence on the time required to establish a telephone connection. The above statements are particularly evident in automatic systems where selector apparatus is required to establish a connection between parties but is released during the conversation. It follows that the number of selector circuits and relays therein depends upon the amount of traffic and time required for the selectors to establish the connection.

To establish a telephone connection between two parties in certain automatic telephone systems, requires the opening and closing of about 2,000 electric switches of which 1,200 are operated by simpler types of electromagnetic relays. In a typical manually operated system a call is completed by the opening and closing of about 112 switches of which 70 are operated by relays. It is therefore evident that the relay switches must operate both quickly and reliably and maintain a high degree of stability throughout a long period of service.

In controlling the various circuits in telephone systems by relays, the character of the circuits determines the construction of the relay switches. If large currents are to be controlled the relay switch

[1] Relays in the Bell System, S. P. Shackleton and H. W. Purcell, *Bell System Tech. Journ.*, Vol. 3, p. 1, 1924.

construction differs materially in ruggedness from the construction where relatively small currents are to be controlled. In the operation of the relays larger amounts of power, of course, are required for those having the more rugged construction. It is also evident that more power is required for fast operation than for comparatively slow operation. Fast operation of relays is also dependent upon

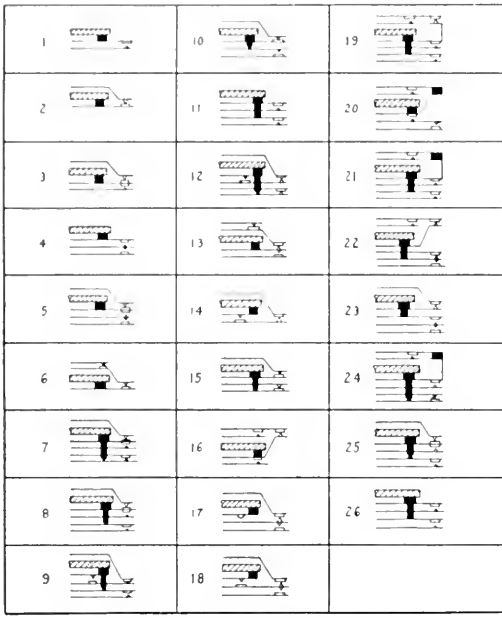


Fig. 1 Spring Combinations Flat Type Relay

circuit arrangements which are effective in lowering the electrical "time constant" of the circuits in which the relays operate.

Electromagnets in telephone systems are designed and used for a great variety of conditions. The more common uses are for relay operation on direct current battery of 20 to 28 volts or 40 to 45 volts. Such relays perform a great number of switching functions, a few of which are shown in Figs. 1 and 12. Other designs are used for oper-

ation on alternating currents, ranging from 16 cycles ringing frequency to voice frequencies of 2,000 cycles per second. The load or work required of these relays and electromagnets varies from a fraction of a gram controlled through a few thousandths of an inch to 25 pounds controlled through a distance of $\frac{1}{4}$ of an inch. Some relays are operated where the annual power charges are negligible while in other designs annual power charges may be controlling. The technical considerations which determine the design features, therefore vary throughout a wide range as to the proportioning of the magnetic parts and the design of the windings. Other general design characteristics that must be carefully considered are as follows:

1. Operating capability of the structure —
   (a) Switching conditions or circuit control required of the relay.
   (b) Design of contacts required to safely carry the energy required by condition (a) throughout the estimated "life" requirements of the switchboard.
   (c) Capability of the structure with respect to the input power to satisfy condition (a).
2. Determination of winding best suited for the circuit.
3. Temperature limitation of the winding under extreme conditions.
4. Ease of adjustment.
5. Permanence of adjustment —
   (a) For a period of service operations representing the "life" of the relay in the switchboard.
   (b) Under extreme weather conditions.
6. Size and mounting facilities.
   (a) When used for additions to old equipment where it should mount in the same space as the apparatus it replaces.
   (b) Economy of space for new equipments.
   (c) Stability of mounting.
7. Terminals — arrangement and distribution for most advantageous electrical connections.
8. Insulating materials.
   (a) Windings.
   (b) Switch control of contacts.
9. Cover design.
   (a) Protection from dust.
   (b) Effect of cover on operation and protection from stray flux.
10. Speed of operation and release.
11. Transmission efficiency with respect to voice frequencies.

12. Mechanical design features with special reference to manufacture.

13. Electro-mechanical efficiency.

14. First cost and annual charges.

As it is not within the scope of the present paper to discuss in detail all of the above characteristics the following have been selected as perhaps the more important and the most interesting:

1. The design of the magnetic parts for various telephone switchboard requirements.

2. Methods of calculating windings and the determination of temperature characteristics.

3. Considerations which determine the spool dimensions.

4. Discussion of designs used extensively in the telephone plant.

Design of Magnetic Parts

The fundamental requirements of an electromagnet or relay are generally the load or pull, the distance through which the load must be moved and the time limits of operation. The last requirement, of course, is reflected in the load or pull requirement as an added pull or force of acceleration.

The fundamental constants of design are the flux leakage coefficient, the core flux density and the flux density in the pole face or area where the pull is exerted. If the designer is given data which fix these constants the remainder of the work is usually a comparatively simple matter of calculation.

The leakage coefficient has been determined experimentally throughout a range of designs where the load to be controlled varied from 1 gram to 5,000 grams. The results show that the leakage depends almost entirely upon the armature air-gap reluctance and the ratio of the core length to the core diameter. The leakage flux is defined as that percentage of the total core flux which does not cross the armature air-gap, and consequently can not be utilized for producing traction. The per cent useful flux is then the ratio of the flux crossing the armature air-gap to the total flux in the relay core. The curves in Fig. 2 for single spool electromagnets and Fig. 3 for double spool electromagnets give the per cent useful flux for various air-gaps and core lengths which are expressed in terms of the core diameter. In cases where the core is round and the pole face area is equal to the

core section these data may be used directly. If, however, the pole face area differs from the core section, the air-gap used in looking up the leakage in Fig. 2 and Fig. 3 should be reduced to a value which,

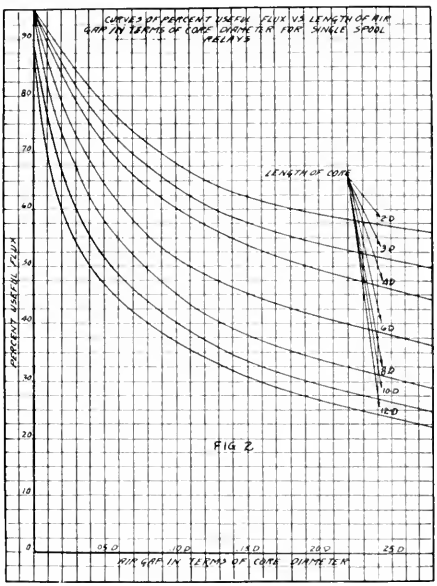


Fig. 2 Curves of Percentage Useful Flux vs. Length of Air Gap in Terms of Core Diameter for Single Spool Relays

with a pole face area equal to the core section, would give the same air-gap reluctance.

The core flux density and the pole face density depend largely upon the requirements of the particular design, but the considerations outlined in the next four paragraphs are of prime importance.

In some cases the annual power charges are relatively unimportant, there being plenty of power available during the short intervals of time required for operation. Obviously in this case efficiency of operation can be sacrificed, and consequently power, in order to

obtain a low first cost. Referring to Maxwell's formula for traction or pull

$$P = \frac{B^2 S}{8\pi 980},$$

the pull P is proportional to the square of the armature air-gap flux density B, consequently the total flux required will be less the greater

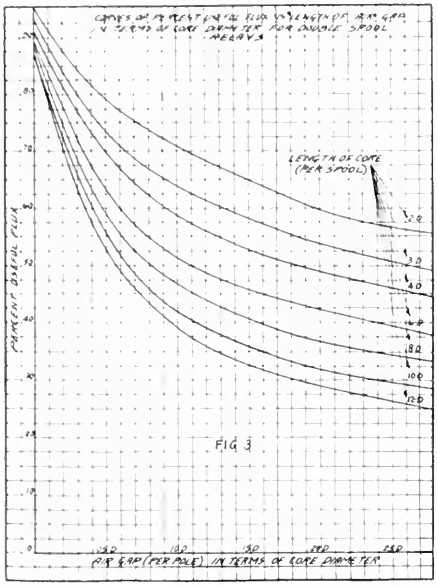


Fig. 3 Curves of Percentage Useful Flux vs. Length of Air Gap in Terms of Core Diameter for Double Spool Relays

the gap density. A high core flux density and pole face density gives a small core section and consequently a small and cheap magnet. The limit to the decrease in size is the allowable temperature limit of the winding.

Of course, there is a limit to the sacrifice of efficiency to obtain a low first cost. If the reasoning in the preceding paragraph is applied to a 5,000 gram electromagnet, the results will show three to

four per cent of the total ampere turns required to saturate the core while on a relay which controls five grams the same assumptions show over 50 per cent of the total ampere-turns required to saturate the core. Where small forces such as five grams are involved, we are almost invariably concerned in maintaining a high efficiency

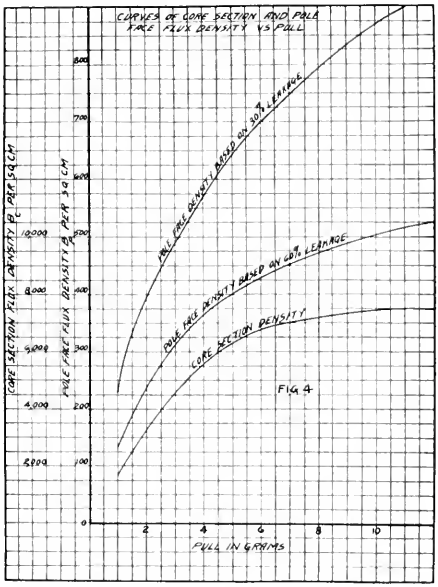


Fig. 4 Curves of Core Section and Pole Face Flux Density vs. Pull

whereas in designs for the heavier forces a few additional ampere-turns required in the core are relatively unimportant.

The work done by an electromagnet is $W = 980\ F\ L$ ergs where F is expressed in grams and L in centimeters. The energy in ergs required to magnetize the core is $W' = \frac{\phi NI}{20}$ where NI represents the ampere-turns required to force the flux ϕ through the core. The ratio of the core energy and the useful work may be taken as a criterion of the efficiency of the core design. Applying this reasoning to

various designs it is found that the most efficient core design is obtained by choosing a core flux density at the maximum permeability of the core iron. If this reasoning is applied to a 5,000 gram relay a saving of approximately five per cent core energy results over working at a high density but the core section is increased in the

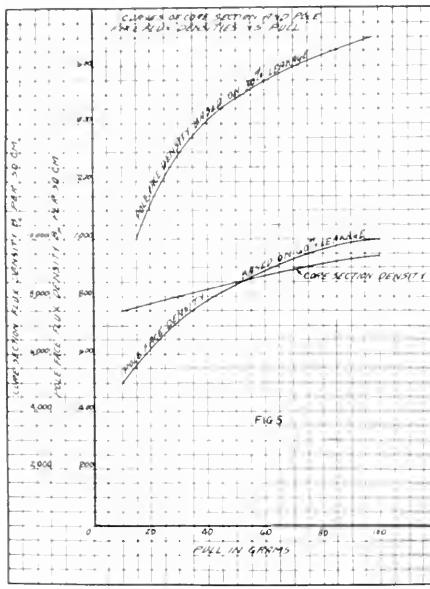


Fig. 5—Curves of Core Section and Pole Face Flux Density vs. Pull

ratio of six to one. Obviously the small improvement in efficiency results in an unreasonable increase in size and consequently first cost and is seldom if ever warranted by the requirements. Applying the same reasoning, however, to a five gram relay we obtain a reduction in core energy of approximately 20 per cent and although the core section has greatly increased this increase has practically no influence on the size or first cost of the magnet. Of course, a further consideration is mechanical strength as where light loads are encountered the core section, needed magnetically, may be entirely too small to

give the requisite mechanical strength for winding or mounting. It may, therefore, be necessary to use a very low flux density in these instances in the core design.

The best flux density and area for the pole face as regards electro-mechanical efficiency is obtained by making the air-gap reluctance

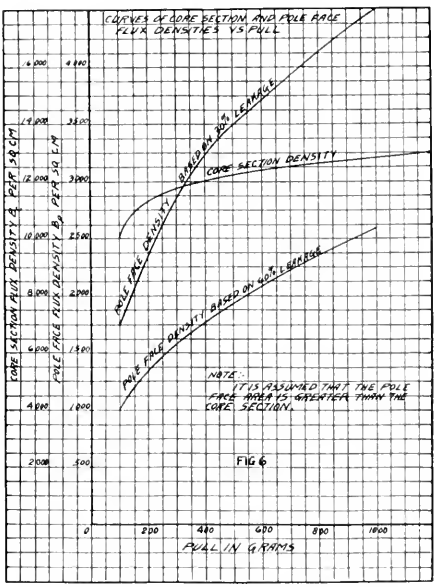


Fig. 6 - Curves of Core Section and Pole Face Flux Density vs. Pull. These Curves Assume That the Pole Face Area is Greater than the Core Section

equal to the reluctance of the remainder of the magnetic circuit. Here again it is found that practical considerations must be carefully weighed, otherwise an unreasonable design results. If, for instance, the pole face density on a 5 gram relay is taken equal to the customary core density, a very small pole face area results. To make the air-gap reluctance, then, equal to the reluctance of the remainder of the magnetic circuit, it is found that an air-gap of possibly .001" or less results. Such a small armature movement, of course, is gener-

ally of no practical value, and consequently, very low pole face densities are generally chosen.

As a result of the above considerations as well as the experience gained in designing a great number and variety of relays and electromagnets, the curves in Figs. 4, 5, 6 and 7 have been drawn which show

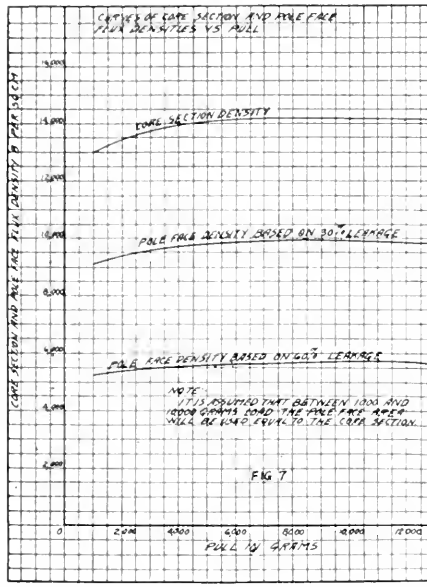


Fig. 7.—Curves of Core Section and Pole Face Flux Density vs. Pull. These Curves Assume That Between Loads of 1,000 and 10,000 Grams the Pole Face Area Will be Used Equal to the Core Section

reasonable assumptions that may be made in working out new designs. These curves are to be employed, of course, with due consideration of the particular requirements in each case.

From the above discussion it is evident that magnetic irons which are capable of high flux densities are particularly desirable for the heavier magnets. The high densities permit of a small core section and consequently a small and low cost magnet. The magnets which control loads of a few grams, however, should be constructed of

magnetic materials which have a high permeability and a low coercive force, but not necessarily capable of working at high densities. A relatively high permeability reduces the energy required to saturate the core although due to the reluctance of the air-gaps there is obviously a limit beyond which no practical gain results due to increased

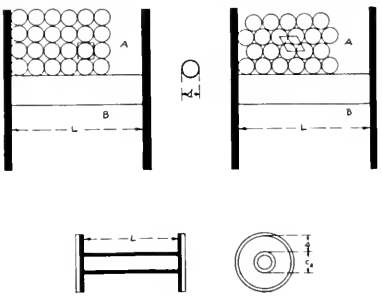


Fig. 8

permeabilities. The most important single requirement of a magnetic material for relays controlling light loads, is a low coercive force. A low coercive force reflects the ability of the magnetic parts to return to practically the same state of magnetization after repeated applications of magnetomotive forces. The effect of residual magnetism, if large, may cause sticking or holding forces of the same order of magnitude as the load requirements. Vacuum annealed silicon steels of comparatively high silicon content and certain nickel steel alloys which have low coercive forces are of great value for electromagnets which must control efficiently light loads of the order of one to fifty grams.

Winding Formulae

Before discussing the economics of the winding dimensions it is necessary to develop and carefully consider the winding formulae and the factors which determine the temperature characteristics.

Fig. 8 shows the one-half cylindrical section of a spool. Since a given wire occupies a similar space in both A and B we need only to consider winding space A. If d in Fig. No. 8 represents the diameter of the wire over the insulation, it is evident that each wire may occupy one of two positions with respect to adjacent wires. In the uniform layup each wire occupies an area d^2, and with the complete inter-

meshing of layers one wire occupies an area $.866\,d^2$. In actual winding practice a combination of the two layups is obtained which gives $.90\,d^2$ square inches as the space occupied by one wire. The area $.90\,d^2$ may be taken as indicating perfect winding so that if the total winding space or area is represented by A and the total turns by N, we have under the best conditions

$$\frac{A}{N} = .90\,d^2.$$

The comparative merit or efficiency of any other winding may therefore be expressed as

$$\frac{.90\,d^2 \times 100}{\frac{A}{N}} = \text{per cent efficiency.}$$

As each size of wire and insulation winds with a somewhat different efficiency, the variation in the value $A \div N$ is generally determined experimentally for each gauge of wire. Thus

$$\frac{A}{N} = K = C_1\,d^2 \qquad (1)$$

The constant C_1 is often designated as a space factor constant and may include the insulating or interleaving paper used throughout the winding. The following are representative values of K for enamel and silk insulated wire of Western Electric Company manufacture.

VALUES OF K

B. & S. Gauge	Enamel Insulated Wire	Silk Insulated Wire
21	.000894	.000936
22	.000718	.000755
23	.000577	.000614
24	.000431	.000477
25	.000437	.0003825
26	.000280	.0003140
27	.000225	.0002615
28	.000183	.0002170
29	.000147	.0001810
30	.000120	.0001510
31	.000096	.0001261
32	.0000781	.0001069
33	.0000628	.0000866
34	.0000516	.0000815
35	.0000410	.0000678
36	.0000338	.0000577
37	.0000269	.0000500
38	.0000222	.0000428

Referring to Fig. 8 the space or area available for winding is

$$A = L\Delta. \tag{2}$$

From equations 1 and 2 the total turns possible are

$$N = \frac{A}{K} = \frac{L\Delta}{K} \tag{3}$$

The total resistance of the winding is the product of the resistance of the mean turn R_m and the total turns N,

$$R_t = R_m N.$$

The length of the mean turn for a round core, Fig. 8, is

$$2\pi\left(\frac{C_2}{2} + \frac{\Delta}{2}\right) = \pi(C_2 + \Delta),$$

and if r is the resistance per unit length we have

$$R_m = \pi(C_2 + \Delta)r,$$

whence the total resistance is

$$R_t = \pi(C_2 + \Delta)rN;$$

or substituting the value of N from equation 3,

$$R_t = \frac{\pi(C_2 + \Delta)r\Delta L}{K}. \tag{4}$$

For a core of rectangular cross section equation 3 holds for the total number of turns and it will be found that the equation for total resistance is

$$R_t = \frac{\pi r L\Delta}{K}\left(\frac{p}{\pi} + \Delta\right), \tag{5}$$

where p represents the periphery of the core in inches.

Temperature Characteristics

The critical circuit conditions with respect to the relay winding specify either constant wattage, constant voltage or constant current. The constant voltage circuit is one in which a change in resistance of the relay winding materially affects the current flow. The constant current circuit is one in which a change in resistance of the relay winding does not materially affect the current flow. An approximate constant wattage condition is one in which a resistance

such as a line in series with the relay is equal to the resistance of the relay and where the resistance external to the relay winding does not change appreciably with temperature variations.

The temperature formulae for the constant wattage condition are developed as follows:

Let Q be the quantity of heat in calories supplied to the winding per second, and $Q\,dt$ be the amount supplied in a small increment of time. Let S be the product of the specific heat and weight of the total wire on the spool expressed in calories. Let T be the temperature difference between the winding and the surrounding air. $S\,d\,T$ is then the amount of heat used in raising the temperature of the wire by the amount $d\,T$. Let ρ be the average dissipating constant throughout the temperature range. It depends upon the radiating surface of the winding, metal conducting parts of the structure and external convection of heat by the air. Given the constant ρ, $\rho\,T\,dt$ represents the calories dissipated during the interval dt.

The total heat supplied during the time dt is partially used in raising the temperature of the wire, and partially dissipated, consequently

$$Qdt = SdT + \rho T dt. \tag{6}$$

If heat is continuously supplied the winding in the form of electrical energy, the rate of dissipation ultimately equals the rate of supply. This is true for temperatures that do not fuse the wire or permanently alter its resistance characteristic. Ultimately

$$SdT = 0$$

and

$$Qdt = \rho T_m dt.$$

If the final temperature reached is designated as T_m then

$$Q = \rho T_m \tag{7}$$

and from equations 6 and 7

$$\rho T_m dt = SdT + \rho T dt,$$

$$-\frac{\rho}{S}dt = -\frac{dT}{T_m - T}$$

and integrating gives

$$-\frac{\rho}{S}t = \log(T_m - T) + C.$$

Observe that when $t = o$ the value of T is also zero and $C = -\log T_m$. Hence

$$T = T_m \left(1 - \epsilon^{-\frac{\rho}{S} t}\right) \tag{8}$$

Equation 8 shows that the transient relation between temperature rise T and time is exponential and ultimately the temperature rise is $T = T_m$.

The final temperature T_m reached by the winding may be determined by writing equation 7 in the form

$$\rho T_m = \frac{EI}{4.186} \tag{21}$$

where $E\ I$ represents the constant wattage applied to the winding and 4.186 is the Joule equivalent. If the room temperature is T_r and the ultimate temperature rise T_m, it is evident that the final temperature of the winding is

$$T_f = T_m + T_r,$$

$$T_f = \frac{EI}{4.186\rho} + T_r.$$

By introducing a new constant K_1 which represents the ability of the structure to dissipate heat and also includes the factor $\frac{1}{4.186}$, we have [2]

$$T_f = \frac{EIK_1}{A_1} + T_r, \tag{9}$$

in which A_1 is the area of the winding but does not include the ends.

The value of K_1 can be readily determined by obtaining an experimental curve between $E\ I$ and T_m. This is obtained by gradually increasing $E\ I$ but holding the wattage constant for each value long enough for the final temperature rise to take place. The value of T_m is calculated by observing the change in resistance of the winding.

The constant current and constant voltage characteristics are determined in a similar manner with the important exception that the quantity of heat Q supplied per second is not constant but varies in accordance with the change in resistance with temperature. Thus for constant current conditions $4.186\ Q\ dt = I^2\ R\ dt$ and for constant voltage conditions $4.186\ Q\ dt = \frac{E^2}{R}\ dt$, where $R = \frac{R_0(234.5 + T)}{234.5}$ for centigrade degrees and R_0 is taken at 0° C.

[2] For single spool relays $K_1 = 50$ to 60, and for double spool relays $K_1 = 35$ to 50.

The steps by which the solution fitting these conditions are obtained will not be given but the results, stated in practical units, are included

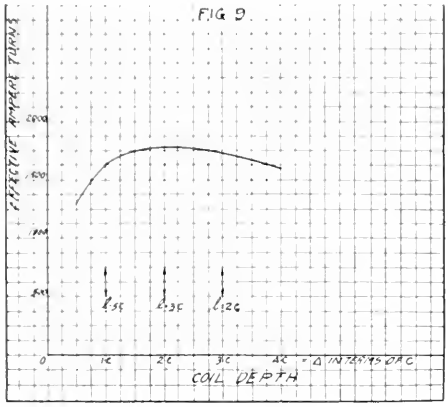


Fig. 9. Relation Between Coil Depth in Terms of Core Diameter and Effective Operating Ampere Turns

to bring out certain important facts relating to winding design. The final temperatures are

$$\text{Constant Wattage, } T_f = \frac{K_1 EI}{A_1} + 20^\circ; \tag{10}$$

$$\text{Constant Current, } T_f = \frac{5090 A_1 + 234.5 I^2 R_{20} K_1}{254.5 A_1 - I^2 R_{20} K_1} \tag{11}$$

$$\text{Constant Voltage, } T_f = -107 + \sqrt{16000 + \frac{254.5 K_1 E^2}{A_1 R_{20}}} \tag{12}$$

The transient temperatures of constant wattage, voltage and current are all of the exponential form $T = T_m (1 - \epsilon^{-t})$, while the cooling of the winding after current is stopped is of the form $T = T_m \epsilon^{-t}$. In these equations ϵ is the constant pertaining to the particular condition considered.

An important observation in connection with these temperature characteristics is the great difference in temperature rise in the three cases with *like initial conditions* of energy input. Thus, it is important to note that an electromagnet which is correctly designed and worked

to its temperature limit in a constant voltage circuit, would overheat in a constant wattage or constant current circuit. A relay properly designed to work at a safe temperature under a constant current condition, would be unnecessarily large and expensive in a constant voltage or constant wattage circuit. It is, therefore, evident that

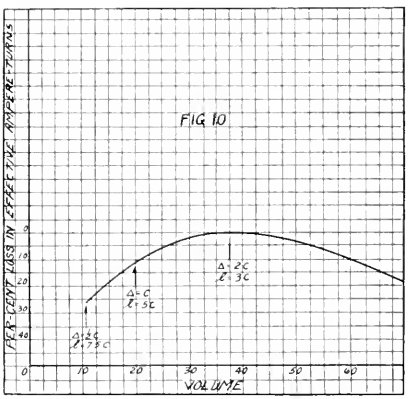


Fig. 10—Relation Between Copper Volume and Percentage Loss in Effective Ampere Terms

exact rules can not be given for the correct proportioning of spool and winding dimensions from a purely design standpoint without consideration of the circuit in which the electromagnet is to operate. Some general design features, however, can be indicated which will enable preliminary assumptions to be made that can be refined as the design is worked out for its particular operating conditions.

Spool Dimensions

Certain important facts regarding spool dimensions are indicated in Fig. 11. The spool dimensions for the winding may be investigated by assuming that a definite radiating surface must be used to dissipate the heat, and then determine the relative values of winding depth, length, and volume in terms of the core diameter. The volume of wire used in the spool is taken as a measure of the first cost and a variation in the length of the coil is reflected in the leakage flux which

in turn may be taken as a measure of the effective ampere turns. The determination of the leakage flux involves reasonable assumptions from experience of the armature air-gap in terms of the core diameter.

If the electromagnet is to be operated on a definite voltage the assumption of a definite radiating surface to dissipate a certain input

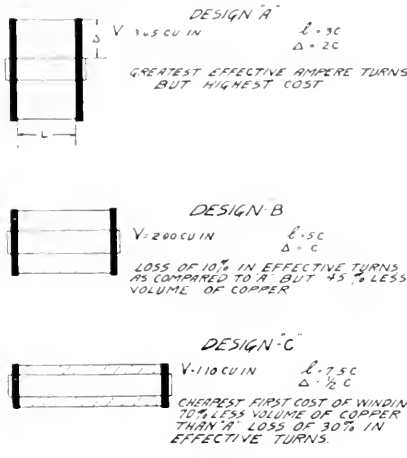


Fig. 11

wattage will fix the resistance of the coil. Copper windings of electromagnets in telephone systems are generally wound with wire which varies from No. 20 B. & S. to No. 39 B. & S. gauge. The resistance generally varies throughout a range of $\frac{1}{2}$ ohm to 2,000 ohms. Various gauges of wire wind with different efficiencies due to variations in the space factor but a number of different gauges may be assumed and the calculations carried out which give the relation between the winding depth and the effective ampere-turns. With a constant radiating surface a variation in the winding depth causes a variation in the length which, of course, is reflected in the leakage flux. The results of a number of calculations on various windings are shown in Fig. 9. In Fig. 10 is shown the relation between the volume of wire on the spool and the per cent loss in efficiency due to a variation in the depth of winding which, with a constant [3] radiating area, causes

[3] The radiating area is taken as the surface only of the coil and the ability to dissipate through the ends and otherwise is reflected by the heating constant K.

a corresponding change in the length of the coil. Fig. 11 shows the relative dimensions of three designs of spools taken from Figs. 9 and 10.

Some very interesting information can be obtained from Fig. 11 in regard to the relation between the volume of wire, as reflecting the first cost, and the ampere turn operating efficiency. Design "A" contains a volume of copper of 3.65 cubic inches, while in design "B" the volume of copper has been reduced to 2.00 cubic inches although the loss in effective ampere-turns is only 10 per cent. In design "C" the volume has been reduced to 1.10 cubic inches with a loss in efficiency of 30 per cent.

Obviously the design "C" is the cheapest in first cost because of the small copper volume and will also give the lowest annual charge where the time of operation is very short and the charge for power relatively low. Where the magnet is required to operate very often and the price of power is high the design "A" will prove the most economical. Design "B" may be considered as intermediate between designs "A" and "C".

In the above considerations of spool dimensions the examples given should not be taken as an accurate generalization but simply as a method which, with a given set of requirements, should enable reasonable first approximations to be made. Thus, if annual power charges are controlling, a relatively short and deep spool will give the best results, although there may be exceptions where for instance, the operating current is reduced to a holding value and where the leakage is relatively small due to the fact that the armature is operated. In such a case and unless operating efficiency is also of prime importance the design "A" would be more expensive than necessary in first cost. Other cases often arise where the input wattage is very small but the operating requirements are very exacting so that the most efficient winding is required and the first cost is relatively unimportant. In this case a larger volume than "A" can be used to advantage. These examples may be used as a guide therefore, in determining spool dimensions which are later refined as the design is completely worked out. The illustrations of designs given in the latter part of this paper show how accurately certain final design dimensions can be worked out to give the minimum annual charge.

Discussion of Designs Used Extensively in the Telephone Plant

To any one familiar with telephone systems it is obvious that it is impracticable to design all the relays required at maximum efficiency and economy for each particular condition that arises. Such a pro-

cedure would involve endless equipment changes as well as the large and unnecessary manufacturing expense of making an excessive number of types of relays. Much of the relay engineering work of the past few years has therefore been directed toward the standardization of relay designs which would be flexible, reliable and economical

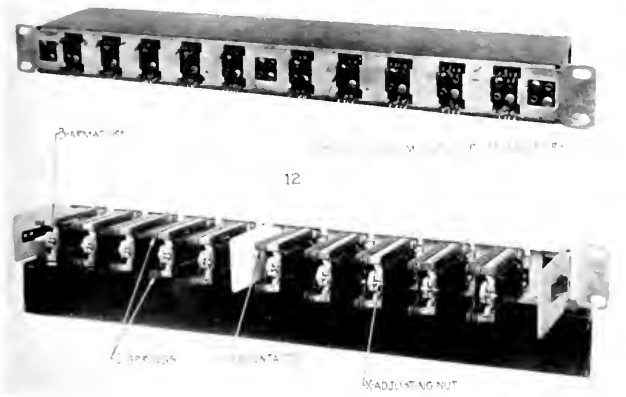


Fig. 12

as a whole in the telephone plant rather than the most efficient in all respects for any specific condition. The flat or punched type relay manufactured by the Western Electric Company represents largely the result of this effort.

The flat relay is essentially a punch press product manufactured yearly in large quantities and in about 3,000 varieties of windings and switching or contacting arrangements. The punch press method produces parts which are exact duplicates and therefore interchangeable which is particularly advantageous both for assembly and replacements or repairs. All the springs as well as the core and armature are punched and formed in bending fixtures to the required shapes. The mounting plates are also punched and designed to permit of uniform and economical mounting of the relays.

A number of these relays are shown on a punched mounting plate in Fig. 12. Referring to the figure it will be seen that the relays are insulated from the mounting plate by phenol fibre insulators "A,"

which are securely fastened to the mounting plate by means of metal eyelets. The armature "B" is hinged at the rear by the use of a thin, steel reed, securely riveted to the armature. The switching arrangements which the armature controls are in the form of nickel silver springs "C" with the contacts "D", at the front and in plain view.

Fig. 13

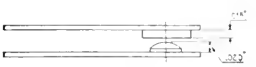


Fig. 14

The springs and contacts are mounted vertically which is particularly effective in keeping the contacts clean. The contact points are made from platinum or a recognized equivalent, and are designed in the form of points and discs to facilitate alignment and adjustment. Two designs of contacts have been standardized; one size being used for the customary electric currents and wear conditions encountered in manually operated systems and a larger size for the somewhat more severe conditions of wear frequently encountered in automatic systems. All contacts are electro-welded on their respective spring supports and the two sizes are shown in Fig. 13 and Fig. 14, respectively.

The springs and their associated contacts are designed in twenty-six switching arrangements as shown in Fig. 1. A single relay may be provided with one of these switching arrangements or any one of these twenty-six arrangements may be paired with any other arrangement. Thus on a single relay there may be chosen any one of 377 switching or contacting combinations. The 377 spring combinations provide a great flexibility in circuit design and permit of uniform and efficient equipment layouts.

In manufacturing the relays the spring assemblies are clamped together under high compression before tightening the screws which hold them together. This insures that the springs retain their position and adjustment throughout a long period of time. The arrangement of the springs is such that definite stops or supports are provided for each spring either on the front spool head or on the armature. In tensioning or adjusting the relay springs against their supports,

sufficient tension is set up in the springs to insure a pressure of at least 15 grams between all contacts at the time of closure.

The amount of current and power required to operate each relay is dependent upon the tension and number of springs that must be moved and the distance through which this movement takes place. Relays or electromagnets operate most efficiently with the armature air-gaps set at the minimum required for the satisfactory opening and closing of the contacts. Consequently a method has been carefully worked out for these relays in which the armature travel is set in accordance with the requirements of the particular spring combination by the adjustment of the friction lock nut "X" shown in Fig. 12. This setting of the armature insures a normal separation of contacts of approximately .010 inch and at least .005 inch "follow" after closure of the contacts. The "follow" allows for a certain amount of contact wear as well as insuring a slight wiping action which gives a certainty of contact closure. The electrical operating current requirements are figured and specified on the basis of obtaining 20 grams pressure between all contacts; this margin being allowed so that no undue hardship will be experienced in maintaining the minimum requirement of 15 grams.

The insulating materials used throughout have been carefully studied and the best materials known to the present day art have been used. Thus the wire used in the winding is insulated with a high grade enamel and the insulating papers on the core are practically inert from an electrolytic corrosion standpoint. The coils are covered with a serving of cotton, treated with unbleached shellac which acts as a seal against moisture and protects the winding from abrasion. The phenol fibre used on the spool heads and spring insulators is much superior to hard rubber in regard to its ability to withstand a wide temperature range without appreciable expansion or contraction.

For this reason it is permissible to work these relays at higher temperatures without danger of fire hazard or deterioration of the insulation than relays insulated with hard rubber parts. These higher temperature limits permit a wider usefulness of the relays in circuits as well as economy in construction as the size of the coil often depends on the necessary area for radiation and this area is fixed by the permissible temperature range.

Where the relays are to remain operated a considerable length of time throughout the day the annual power charges become important and the design of the winding and in some cases the size of the spool must be altered to give the minimum annual charge. The group of

curves in Fig. 15 show how nearly correct these relays have been designed for conditions where the operating ampere-turns are 260 and the relays remain operated from 60 to 600 minutes per day.

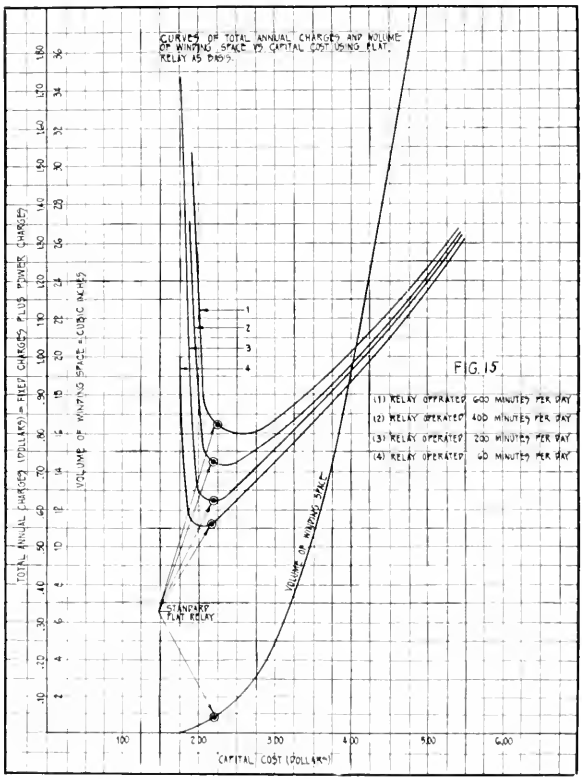


Fig. 15

The capital cost and annual charge figures should be taken as relative only as the correct values will vary with manufacturing conditions and with the cost of power for different localities.

Other designs of relays used extensively in the telephone plant are the relays that control the supervision of a telephone connection and the alternating current relays which operate on ringing currents of 16 to 20 cycles frequency.

Relays which are used for supervisory purposes and alternating current relays are generally constructed of silicon steel instead of the customary Norway or magnetic iron. The silicon steel is very

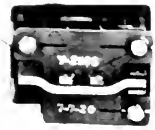


Fig. 16

satisfactory for these relays because of its comparatively high permeability, low coercive force and small hysteresis. The high permeability is advantageous for relays that are required to operate on a very small energy input and the low coercive force is very effective for obtaining a quick and positive release of the relay armature, particularly where a leak current exists due to faulty line insulation. A great improvement in many of these relays can be obtained by the use of certain nickel-iron alloys which have been recently developed and are known as "Permalloy."

A relay for use on ringing currents is shown in Fig. 16. The armature "A" of this relay is attracted to the bifurcated extensions of the core "B." One of these core extensions is completely surrounded by a part of the copper spool head "C." This arrangement is known as pole "shading" or phase splitting and is used to produce a substantially steady pull on the armature when the relay is energized by single phase alternating current.

Referring to Fig. 17 the theory of operation is shown by considering the vector diagram in connection with the schematic drawing of the relay core and armature. When an alternating current is applied to the winding we can assume that an alternating flux $2\,\phi_m$ is generated in the core. This flux divides into two approximately equal parts in the two bifurcated extensions of the core. If these two fluxes can be displaced in time phase it is evident that the armature will be attracted by one of the bifurcated extensions of the core, while the flux, and consequently the attraction of the other, is passing

through zero. This may be explained by the vector diagram in which E_2 represents the induced voltage in the short circuited copper ring due to the alternating flux ϕ_m. The current in the copper ring I_2 lags behind the voltage E_2 as shown and the flux due to this current

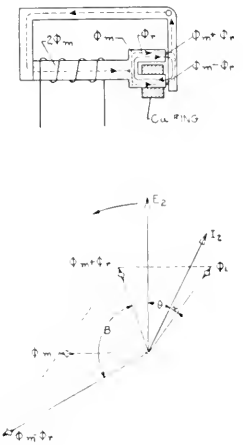


Fig. 17

is ϕ_r. This flux ϕ_r has a magnetic path through the bifurcated pole pieces and armature as shown by the arrows. Following out the arrows it will be seen that this flux adds to the flux ϕ_m in the upper part and subtracts in the lower part of the two core extensions.

The vector addition and subtraction of these two fluxes results in two vectors $\phi_m+\phi_r$ and $\phi_m-\phi_r$, each of which represents a flux that crosses an air-gap to attract the armature. These two fluxes differ in time phase as represented by the angle "B" so that a substantially constant attraction results on the armature. The operation of the relay under these conditions is very much the same as that of a direct current relay as no vibration or chatter of the armature or contacts occurs. The minimum effective alternating current ampere-turns required for operation are 70 to 100 ampere-turns.

Such a relay, of course, operates on direct current as well as on alternating current and in fact the direct current supervisory relays are quite similar to these relays in mechanical design.

Fig. 18 shows the design features for the supervisory and ringing frequency relays. In this figure the winding has been omitted so as to show clearly the unusually small core. This construction is especially efficient in circuits where the relay receives at times a

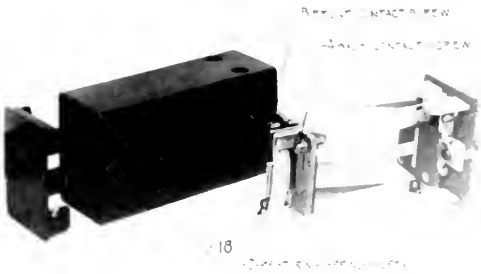


Fig. 18

very small amount of energy for operation and must also release reliably against a leak current immediately after operation by a comparatively large amount of energy. The small core saturates magnetically on a relatively small current or energy so that excessive energy does not store up additional magnetism which would retard or prevent the release of the relay.

Referring further to Fig. 18 the micrometer screws "A" and "B" are used to adjust the back and front contacts respectively, and to fix both the unoperated and operated positions of the armature. The screw "C" is used to control an armature restoring spring which is in the form of a flat spring riveted to the armature. These relays are generally provided with individual covers which are effective in preventing cross talk of telephone voice frequencies when used as supervisory relays in telephone switchboards.

A Dynamical Study of the Vowel Sounds

By I. B. CRANDALL and C. F. SACIA

Introduction

THE study of the vowel sounds presents a problem which has interested scientists and scholars in varied fields. A knowledge of their nature is of fundamental importance not only in communication engineering but also in acoustic science, phonetics and vocal music. From the earliest theories and the rough experiments of Willis (1829) and Helmholtz (1859) to the later measurements of D. C. Miller (1916) steady progress has been made toward the accurate determination of their characteristics.

Further progress in this study has been made possible with improved facilities now available in the telephone research laboratory. It has been felt that there was need for more accurate records of the spoken sounds and the development of improved transmitters, amplifiers and other devices has made possible recording apparatus of greater accuracy, range and power than any heretofore used.

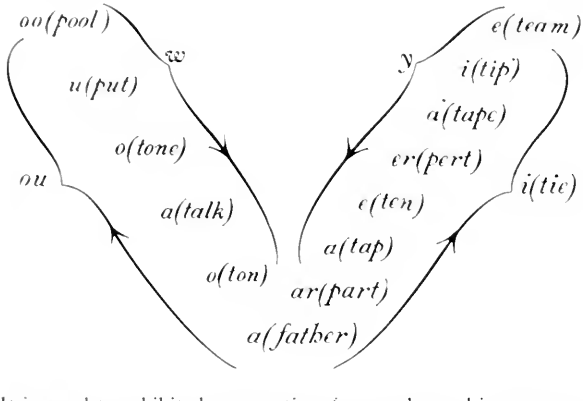
In this paper will be given the results of an analysis of spoken vowel sounds based on a set of accurate oscillographic records. The recording apparatus was designed to record the wave forms of the different speech sounds practically free from distortion over the frequency range from 100 to 5000 cycles. A brief description of this apparatus is given in the appendix. The emphasis in the present paper is placed on the composite frequency characteristics of the sounds as revealed by a particular method of analyzing the records so obtained.

Analysis of the Data

The thirteen vowel sounds investigated are shown arranged in a triangle in Fig. 1. The diphthongs *ou*, *w*, *y* and long *i* are not included. Eight records of each sound were taken, four by male and four by female speakers. In speaking these sounds the only constraint imposed on the speakers was that the sound should be completely uttered within an interval of one second. The recording mechanism was so arranged that the whole of the sound from beginning to end was recorded in one continuous graph. In practice the average duration of these sounds was about 0.30 second. Each record shows a sequence of growth and decay in amplitude somewhat as follows: first a period of rapid growth in amplitude lasting about .01 second during which all components are quickly produced

and rise nearly to maximum amplitude; second a middle period in which the general amplitude is nearly constant but with varying phase relations between the different components and lasting about 0.17 second; and finally a period of gradual decay lasting about .09 second in which all the components disappear. A typical record so obtained is shown in Fig. 2.

A brief description of the method of mechanically analyzing such a record is given in the appendix. The essential point of the analysis is that the whole record from start to finish is taken as the unit for analysis and the data obtained are therefore the average characteristics of the sounds throughout their duration.



It is usual to exhibit the properties of a vowel sound in a spectrum diagram showing the amplitude of the component vibrations as a function of their pitches or frequencies. For each vowel sound there are, in addition to fundamental tones, certain characteristic regions of resonance which may be at high or low frequencies. It would be possible from the results of this analysis to present the sound spectra of each vowel showing the relative amplitudes for the different frequencies as present in the original air vibration[1] but this treatment has been modified to take into account the relative importance of the various pitches in hearing. Using the data available

[1] In previous publications (*Phys. Rev.* XIX, 1922, p. 228, Fig. 7, and *Bell System Technical Journal*, Vol. 1, No. 1, p. 121,) data have been given showing the actual distribution of energy in average speech. The tremendous concentration of energy in the lower frequencies is somewhat misleading unless account is also taken of the much reduced sensitivity of the ear in this region.

on the relative sensitivity of the ear at different frequencies[2] we have multiplied the acoustic amplitude at each frequency by the corresponding ear sensitivity factor and the results obtained are taken to be the effective amplitude frequency relations which are characteristic of these sounds.

The data from the four male records and from the four female records of each sound are separately composited and the resulting curves are shown in the diagram (Fig. 3). This compositing process was somewhat laborious because the analyses of the separate records were made not with reference to predetermined frequency settings, but rather for those critical frequencies which best determined the shapes of the spectrum curves. The individual curves were therefore plotted, and the average ordinates were then read off for small intervals of pitch. These ordinates were then averaged for each group of four analyses. These average ordinates (after being corrected for the calibration of the recording apparatus) were then multiplied by the ear sensitivity factors for the corresponding frequencies, and the curves so obtained were plotted on the musical pitch scale according to the usual practice. The final spectrum diagram thus shows the relative importance of the amplitudes of all the components of each vowel for male and female speakers.

The amplitude units are entirely arbitrary; it is only the shapes, not the sizes of these curves which have any significance. The order in which these curves are arranged is based upon the vowel triangle in Fig. 1.

Characteristics of the Vowel Sounds

The results of the analyses, as given in Fig. 3 show the essential dynamical properties of these sounds. Consider first the sounds numbered I to VI, which include those vowels usually designated as having single regions of resonance. Progressing through the sequence from I to VI this region of resonance rises in average frequency and becomes narrower in range. The rise in average frequency is of course a well known characteristic. There is also, at least with the male voices, a somewhat scattered and less well defined high frequency range of resonance, perhaps not essential in speech but more highly developed in well-trained singing voices.

The sound *a* (No. VI) is as it were the center of gravity of the vowel diagram and occupies the key position in the phonetics of

[2] See this Journal Vol. II, No. 4, October, 1923. The paper on audition, by H. Fletcher shows a cut of the "Threshold of Audibility" curve from which these data were obtained.

most languages. Now consider the sequence from this sound to No. XIII at the end of the diagram; these sounds include most of those which are known to have two characteristic regions of resonance. The main region of resonance now divides into two parts which gradually recede from each other as we follow the diagram downwards. (Sound X (*er*) is difficult to fit into the diagram in an exact position, but it is evident that it belongs in the series of doubly-resonant vowels.)

Contour lines (nearly vertical) have been drawn on the diagram to indicate the progressive changes in regions of resonance. Viewing the diagram as a whole it is important to consider not only the location of the resonant ranges but also their extent, and their relative separation from other resonant ranges in order to arrive at the essential characteristics of the vowel sound. In other words the individual vowel characteristic depends not only on the absolute pitch but on the relative pitches in case there is more than one region of resonance. It is only in this way that we can explain what is a matter of universal experience in using the phonograph; namely that moderate variations from normal speed in recording and reproducing speech leave the vowel sounds still intelligible.

It is expected to deal in a later publication with the semi-vowel sounds *l*, *ng*, *n*, *m* which seem to be related to the general diagram of the vowel sounds, and on which a preliminary report has already been made[3].

The more interesting features of the original records as such will also be dealt with in a subsequent publication.

APPENDIX

Recording and Analysis of Vowel Sounds

Recording Apparatus

The apparatus used in recording consisted of a condenser transmitter, an amplifier, and an oscillograph, in which important modifications were made. The vibrator was given great stiffness and damping so that the frequency response of the vibrator was nearly uniform up to 5000 cycles. Instead of the usual 12 inch film, special film 51 inches in length was used. This necessitated a much larger film drum. Furthermore the desired length of the record was about four times the circumference of the film drum, so the shutter was arranged to stay open during four revolutions while the vibrator was

[3] *Phys. Rev.* 23, 1924, p. 309—"Preliminary Analysis of Four Semi-Vowel Sounds."

given a slow uniform rotation about its vertical axis. With the film on the drum, the record thus had a helical form. In this way records of the requisite length were obtained.

The condenser transmitter was of the type developed by E. C. Wente, its characteristics combining with those of the amplifier and oscillograph vibrator in such a way that the combined amplitude response for the whole system was fairly uniform up to 5000 cycles, while the phase lag was approximately a linear function of frequency over the same range. This apparatus was therefore well adapted to the production of faithful records of the vowel sounds. The photographic equipment permitted the use of a time scale as great as six meters per second on the record (i.e. 2 inches = 0.01 sec.)

Transformation of Records for Analysis[4]

The oscillograms taken with the above apparatus were line records; in order to analyze these wave forms by the photo-mechanical method outlined below, it was necessary to transform the line record into a black profile. This was accomplished in the following steps:

(1) A positive print of the wave form on the original record was made on motion picture film.

(2) The emulsion of the positive print was then cut through to the base along the line of the wave by means of a stylus.

(3) The entire strip was blackened (on the emulsion side) with printer's ink.

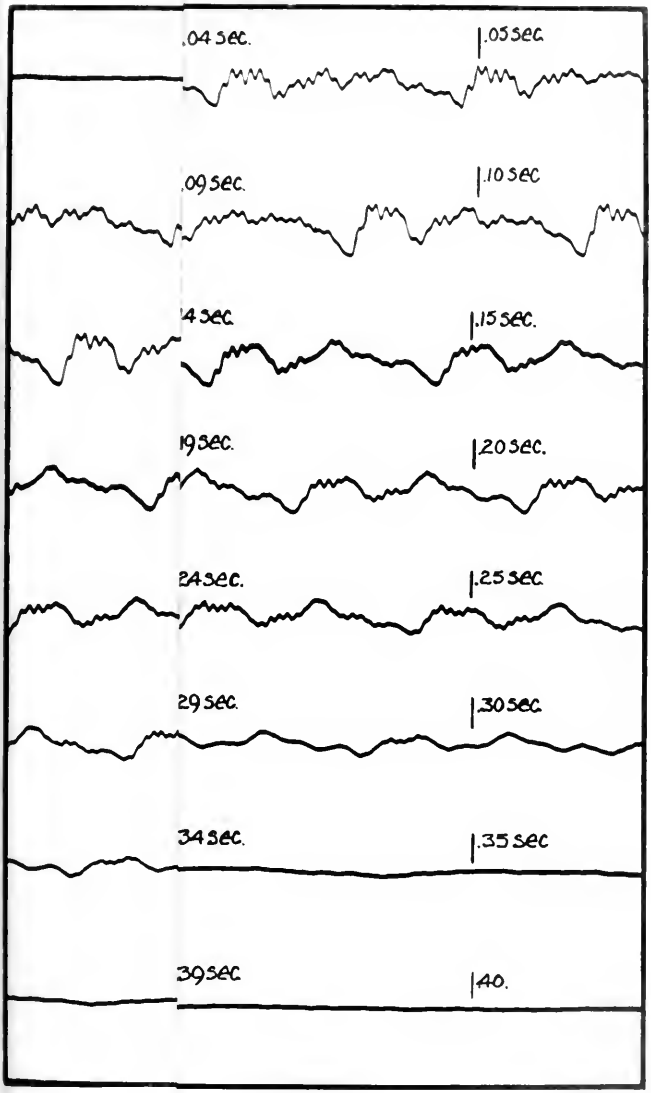
(4) The emulsion on one side of the wave was stripped from the base, thus leaving the profile.

(5) The beginning and end were joined to form an endless belt.

Photo-Mechanical Analysis of the Prepared Records[4]

The principle of the photo-mechanical analysis is as follows: The motion of the strip past the image of an illuminated slit causes fluctuations in a beam of transmitted light which in turn, produce voltage fluctuations in the circuit containing a selenium or photo-electric cell. This voltage is then analyzed by means of a tuned circuit, an amplifier and a rectifier. The frequency of any component selected in this manner is determined by the tuning frequency divided by the ratio of speed transformation (analysis speed divided by the original speed of recording). The measured amplitude of the selected

[4] *Phys. Rev.* 23, 1924, p. 309. It is planned to publish a more detailed description of this apparatus later.


.04 sec.
.05 sec.
.09 sec.
.10 sec
.14 sec.
.15 sec.
.19 sec.
.20 sec.
.24 sec.
.25 sec.
.29 sec.
.30 sec.
.34 sec.
.35 sec
.39 sec.
.40.

given a slow uniform rotation about its vertical axis. With the film on the drum, the record thus had a helical form. In this way records of the requisite length were obtained.

The condenser transmitter was of the type developed by E. C. Wente, its characteristics combining with those of the amplifier and oscillograph vibrator in such a way that the combined amplitude response for the whole system was fairly uniform up to 5000 cycles, while the phase lag was approximately a linear function of frequency over the same range. This apparatus was therefore well adapted to the production of faithful records of the vowel sounds. The photographic equipment permitted the use of a time scale as great as six meters per second on the record (i.e. 2 inches = 0.01 sec.)

Transformation of Records for Analysis[4]

The oscillograms taken with the above apparatus were line records; in order to analyze these wave forms by the photo-mechanical method outlined below, it was necessary to transform the line record into a black profile. This was accomplished in the following steps:

(1) A positive print of the wave form on the original record was made on motion picture film.

(2) The emulsion of the positive print was then cut through to the base along the line of the wave by means of a stylus.

(3) The entire strip was blackened (on the emulsion side) with printer's ink.

(4) The emulsion on one side of the wave was stripped from the base, thus leaving the profile.

(5) The beginning and end were joined to form an endless belt.

Photo-Mechanical Analysis of the Prepared Records[4]

The principle of the photo-mechanical analysis is as follows: The motion of the strip past the image of an illuminated slit causes fluctuations in a beam of transmitted light which in turn, produce voltage fluctuations in the circuit containing a selenium or photo-electric cell. This voltage is then analyzed by means of a tuned circuit, an amplifier and a rectifier. The frequency of any component selected in this manner is determined by the tuning frequency divided by the ratio of speed transformation (analysis speed divided by the original speed of recording). The measured amplitude of the selected

[4] *Phys. Rev.* 23, 1924, p. 309. It is planned to publish a more detailed description of this apparatus later.

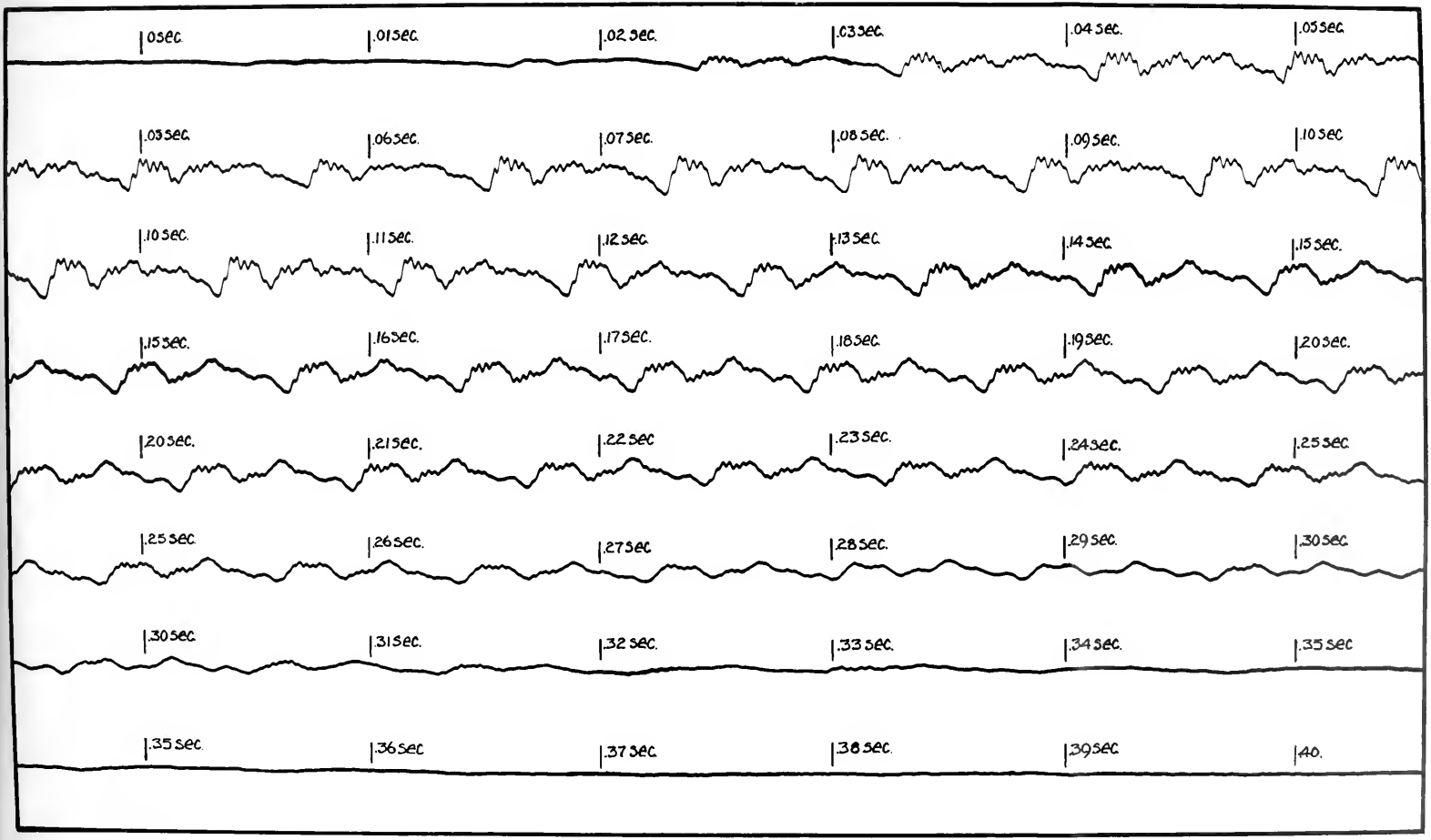


Fig. 2—*oo* as in *pool*; spoken by M. B.—male, low pitched. Plate No. 2; made from film record No. 157-A.

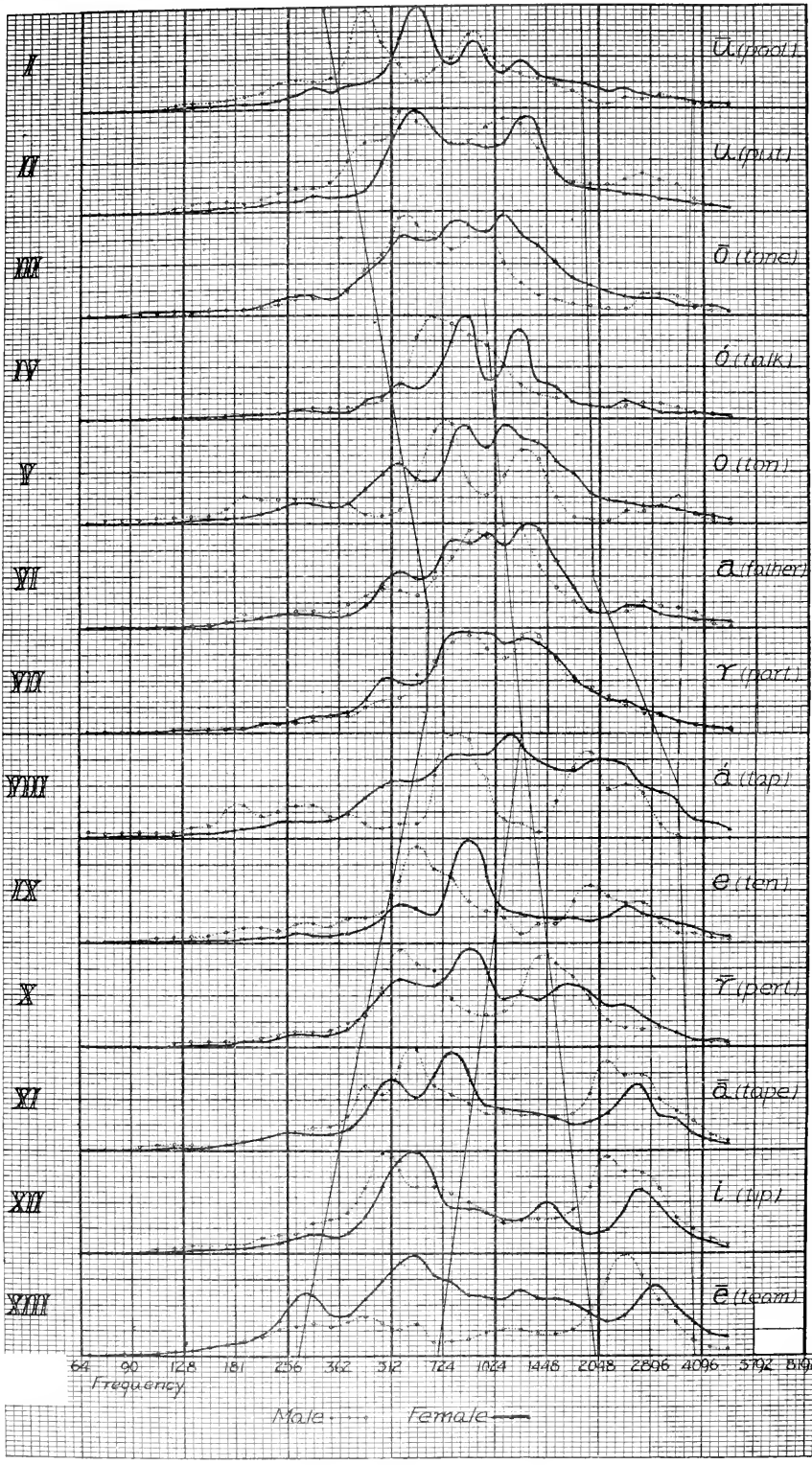

I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII
U (pool)
U (put)
O (tone)
Ó (talk)
O (ton)
a (father)
Y (port)
á (tap)
e (ten)
Ȳ (pert)
ā (tape)
i (tip)
ē (team)
64 90 128 181 256 362 512 724 1024 1448 2048 2896 4096 5792 8192
Frequency
Male
Female

Fig. 3

component is determined by the rectifier output, the sensitivity factor of the selenium cell and the area of the frequency response curve of the tuning apparatus.

Since the wave form of a vowel sound is not a true periodic function, it is represented analytically by a Fourier Integral, not by a Fourier Series. The continued repetition of the motion of the wave past the slit, however, builds up a periodic function consisting of a fundamental and a series of harmonics. The magnitudes of these components bear a simple relation to those of the infinitesimal components of corresponding frequencies in the Fourier Integral. It is this series of harmonics which is measured by the above method, hence the problem of analyzing the aperiodic function represented in the record is solved by means of the related periodic function.

Humidity Recorders

By E. B. WHEELER

DURING recent years, the study of atmospheric conditions and their bearing on various industrial problems from the standpoint both of their effects on human efficiency and on manufacturing processes, is a matter that has received much attention, and the use of air conditioning systems, which have been developed in the last few years, has resulted in greatly improved working conditions, as well as in increased outputs of manufactured products of better quality than obtainable when air conditioning was not employed.

It is not so well appreciated, perhaps, that atmospheric conditions have a material effect upon the operation of intricate electrical and mechanical apparatus, such as those found in telephone systems.

Water vapor, and both gaseous and solid impurities in the air, hasten oxidation and corrosion of metals and also reduce the value of the insulation afforded by insulating materials. These effects usually are greatly accelerated if the temperature is high and if the materials are subjected to differences of electrical potential. Telephone apparatus and equipment consist of combinations of materials which are subject to both of these effects and, in general, the parts are small and the materials used in making them must be carefully chosen with regard to the necessary physical and electrical properties required for proper functioning of the apparatus. Therefore, the severe atmospheric conditions, which may be encountered in service, either must be eliminated by the use of air conditioning systems or the apparatus must be designed to withstand those conditions.

Accordingly, in order that the problem may be handled intelligently, accurate information must be available showing the character of the atmospheric conditions which exist in typical localities where telephone equipment is installed, so that the effects of these conditions on proposed designs may be studied under carefully controlled similar conditions in laboratory "humidity rooms." An outline of some of the work which has been done in an effort to obtain such information may therefore be of interest.

The first recourse would seem to be the data recorded by the various stations of the United States Weather Bureau. However, since these data usually represent periodic observations of outdoor conditions which are obtained primarily for meteorological purposes, it was found that while they indicate the general climatic conditions of different localities, they can not be taken to represent typical conditions in central office buildings, and therefore it has been

necessary to devise methods by which we might secure such information.

The subject of hygrometry has long been one of the problems to which various investigators have given attention and the results of their work are a matter of record.

Thus it has been recognized [1] that, because of its ease of manipulation and its accuracy if suitable precautions are observed, the ventilated psychrometer is a suitable instrument for use in humidity measurements.

Consideration of the various types of hygrometers, commercially available, indicated however, that none would be suitable if reliable continuous records were to be secured. The use of simple wet bulb —dry bulb hygrometers would require practically constant attendance if frequent observations were made, and the results would not be accurate unless arrangements were made to circulate the air over the wet bulb. A pen recorder of the circular chart type to record wet and dry bulb temperatures had been used during one summer in a telephone central office where the humidity conditions were severe, but the results secured were not considered reliable because of the unsatisfactory method used to ventilate the wet bulb, as well as the sluggishness of the recorder due to pen friction on the chart.

Considerable experience in the laboratory with a recording hair hygrometer also had shown that, in addition to the inaccuracies to which hair hygrometers are commonly subject, the friction in the lever mechanism and between the pen and the chart made the instrument too erratic to be considered of possible use in the work being undertaken. Accordingly, a study was made to determine the possibility of developing apparatus which would overcome the troubles inherent in such recorders.

Development of a Recording Hygrometer

A promising method, developed by D. T. May of the Bell System Laboratories and operated successfully in the laboratory, consisted in the use of accurate and matched mercury thermometers, the stems of which were contained in a camera which would enable the heights of the mercury columns to be photographed upon a roll of sensitized paper. Arrangements were made for shifting the paper between

[1] U. S. Weather Bureau Psychrometric Tables for Obtaining the Vapor-Pressure, Relative Humidity and Temperature of the Dew-Point from Readings of the Wet and Dry Bulb Thermometers, by C. S. Marvin.

Proceedings of the Physical Society of London, Feb. 15, 1922. The Measurement of Atmospheric Humidity, by Sir Napier Shaw.

exposures, and a small exhaust blower was provided for circulating the air over the wet bulb. The whole apparatus was controlled electrically by a clock and was arranged to record the wet and dry bulb temperatures at any desired time interval. When the complete record roll had been exposed it was removed and upon development showed the thermometer readings from which the corresponding humidities could be found in the psychrometric tables. While this type of recorder would no doubt have enabled accurate information to be obtained, it had two inherent objections. These were, first, the bulkiness of the complete equipment which had to be placed at the location where the conditions were to be determined and,

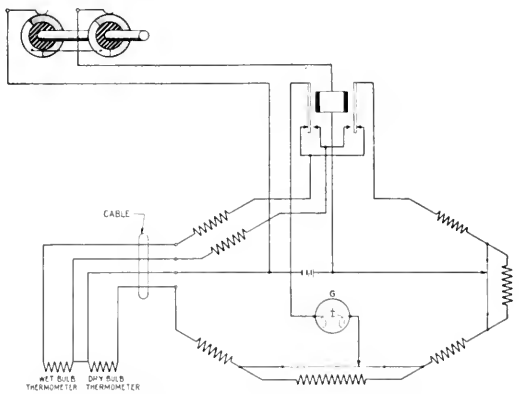


Fig. 1—Bridge Circuit of Difference Recorder

second, the thermometers could not be read because their stems were within the camera box, and therefore, the humidities and temperatures measured could not be ascertained until the record had been developed.

Accordingly, at this time, consideration was given to a type of mechanism which would produce a visible record upon a chart continuously available for observation by the operator. It was found that the Leeds & Northrup automatic recorder had been in commercial use for some time for the measurement of furnace temperatures, by means of thermocouples in conjunction with an automatically adjusted potentiometer circuit. The same type of recorder also had

Fig. 2—Wind Tunnel Equipment

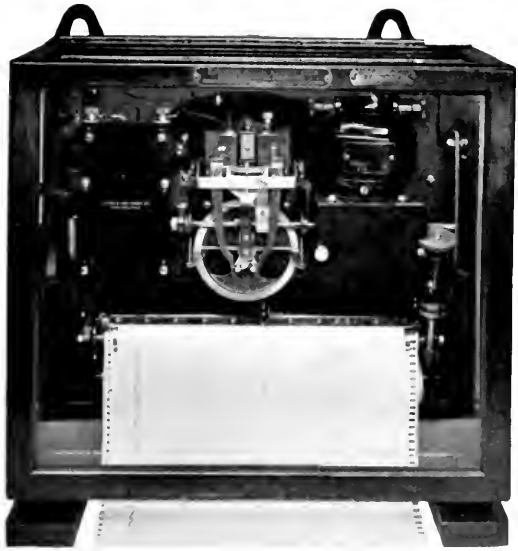


Fig. 3—Temperature and Difference Recorder

been used for recording temperatures and differences between two temperatures by means of resistance thermometers and a Wheatstone bridge arrangement. As it seemed feasible to adapt this instrument to meet our requirements, the double Wheatstone bridge circuit shown in Fig. 1 and the auxiliary wind tunnel equipment with resistance thermometers shown in Fig. 2 were developed. Fig. 3 is an illustration of the Leeds & Northrup recorder used.

This recorder was arranged to measure the resistance of the dry bulb thermometer and the difference between the resistances of the dry and wet thermometers, and to record these values upon a chart. Referring to the circuit diagram Fig. 1, it may be seen that, by means of a relay whose operation is controlled by the commutator on the recorder mechanism, the two Wheatstone bridges, one containing the dry bulb thermometer, and the other containing both the dry and wet bulb thermometers, may be balanced alternately by the recorder. After a sufficient interval has elapsed in each case for the bridge to become balanced the siphon pen is lowered into contact with the chart by a cam mechanism and the point of balance thus recorded. The record thus produced consists of dotted curves showing the successive indications of dry bulb temperature and difference between dry and wet bulb temperatures.

In order to secure the desired accuracy and sufficient sensitivity to follow the changes in temperature, the resistance thermometers used consist of platinum wire wound on mica cards and encased in flat nickel silver tubes with hard rubber ferrules. These are attached to a brass junction box in which is terminated the four conductor cable leading to the recorder mechanism.

The thermometers are enclosed in slotted brass tubes through which the air is drawn by a small blower driven by a universal motor. Mounted below these tubes is a shallow, covered water tank having a slot in the cover beneath the wet bulb thermometer through which the wick projects into the water. The desired water level in the tank is secured by an inverted water bottle, the neck of which projects into another opening in the cover of the tank.

The wind tunnel equipment[2] containing the resistance thermometers may be placed at any desired distance from the recorder mechanism, as the resistances of the thermometer leads have no effect upon the measurements provided they are equal. Leads consisting of a four conductor rubber insulated lead covered cable from 50 feet to 100 feet in length have been used.

[2]The wind tunnel and equipment is quite similar in operation to the "distance hygrometer," *Sci. Am.* June 6, 1914, p. 468.

As one of the difficulties encountered in the use of the wet bulb thermometer consists in the gradual clogging and drying up of the wick due to the accumulation of impurities left in it from the evaporation of the water, together with the dust which settles from the air which is drawn over it, special care must be taken to guard against trouble from this source. The cotton fabric used for the wicks which cover

Fig. 4.—Ventilated Psychrometer

the wet bulb must be treated to remove all traces of grease, with subsequent thorough washing to remove all traces of corrosive material. After this, the wicks should be handled only with thoroughly cleaned hands before they are placed on the thermometers. These wicks should be changed daily. Pure distilled water must be used in the tanks and they must be cleansed occasionally because they become contaminated by the impurities washed out of the air as it bubbles

through the water. By rigid observance of such precautions no difficulty should be experienced in securing accurate records by means of this recorder.

Laboratory Tests

Several of these recorder mechanisms were built and after having been adjusted to operate satisfactorily, each wind tunnel equipment connected to its associated recorder was placed in a laboratory room controlled by air conditioning equipment, and given a run to test its operation under the range of conditions which might be expected to occur at the localities where the recorders were to be installed. During this test, the readings given by the recorder were compared with those obtained with a ventilated psychrometer, Fig. 4, equipped with accurate wet and dry bulb thermometers. Table I following gives a summary of the readings obtained in calibrating one of the recorders, while Fig. 5 shows a typical 12 hour record obtained in one of the laboratory rooms.

TABLE I

Ventilated Psychrometer			Leeds & Northrup Recorder			Per Cent Difference
Dry Bulb Temp. F	Difference between Dry and Wet Bulb, Temp. F°	Relative Humidity Per Cent	Dry Bulb Temp. F°	Difference between Dry and Wet Bulb, Temp. F°	Relative Humidity Per Cent	
77.9	11.5	54.0	77.8	11.1	55.5	+2.8
77.2	10.2	58.0	77.0	9.9	59.5	+2.6
78.4	0.9	96.5	78.2	0.8	97.0	+0.5
84.4	0.7	97.0	84.1	0.6	97.5	+0.5
83.5	5.6	77.5	83.6	5.6	77.5	0.0
83.7	10.8	59.5	83.4	10.5	60.5	+0.1
98.3	10.5	65.5	98.2	10.5	65.5	0.0
98.3	13.4	57.0	98.3	13.4	57.0	0.0
97.4	1.3	95.0	97.8	1.5	94.5	−0.5
97.2	1.0	96.0	97.6	1.2	95.5	−0.5

Reference to these tabulated values of relative humidities obtained by the two methods indicates that the recorder is capable of giving reliable data particularly through the range of high humidities where the effects on materials or apparatus exposed to these conditions may be large. Difficulty was experienced in comparing the readings of the two instruments due to the sensitivity of the resistance thermometers to slight temperature changes, and also due to the slight differences in temperature between the two sets of thermometers which necessarily occurred because they were not in the same wind tunnel. This difficulty was encountered particularly when the "humidity

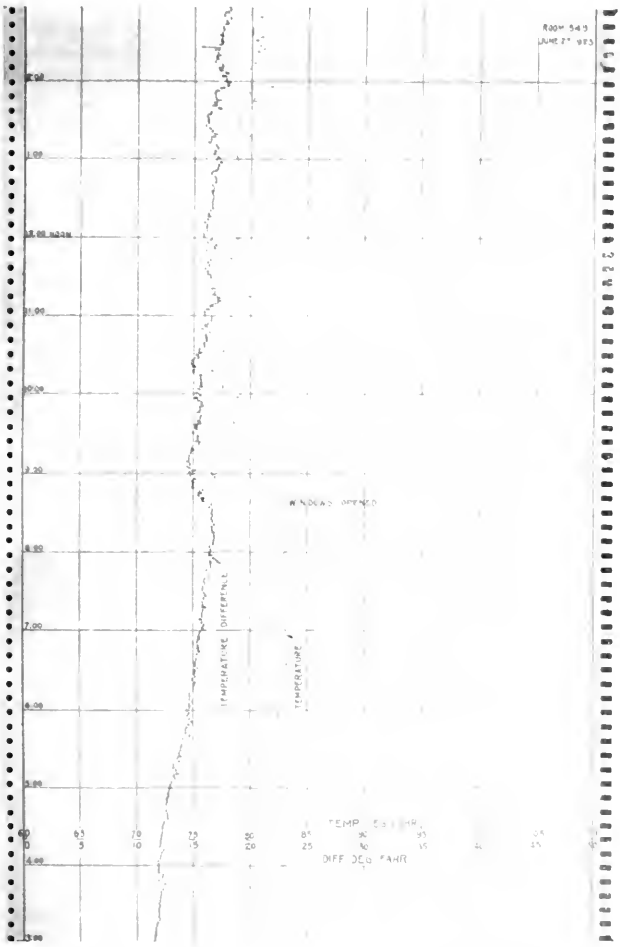


Fig. 5.—Record from Temperature and Difference Recorder

room," in which the apparatus was located, was under a thermostatic control which allowed a temperature variation of approximately $\pm 0.5°$ F. However, the calibration of the resistance thermometers and the sensitivity of the bridges in which they are placed is such that temperatures and temperature differences are recorded with an accuracy of $\pm \frac{1}{4}°$ F.

Field Trials

In order to determine just what combinations of temperature and relative humidity prevail in widely separated localities of the United States, certain cities were selected in which moisture troubles with

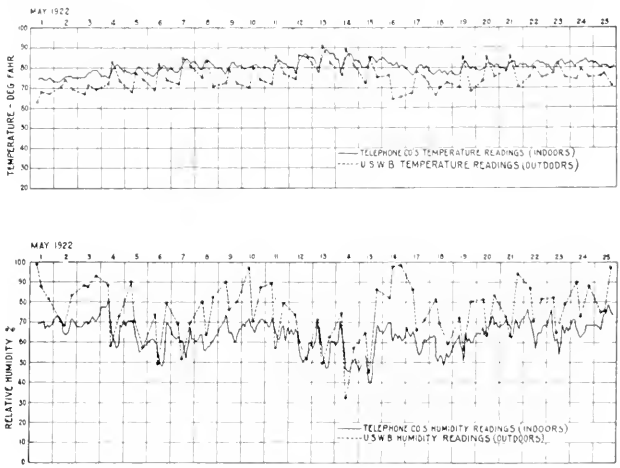


Fig. 6 Comparison of Indoor and Outdoor Temperatures and Relative Humidities at Savannah, Ga.

telephone equipment might be expected to occur, and at which local stations of the United States Weather Bureau were located, so that comparisons might be made between our records of indoor conditions and the observations of outdoor conditions.

Ten of these instruments were installed in central offices in New York (3), Boston, Savannah, New Orleans, Chicago, Minneapolis, Houston and Seattle, from which records have been obtained during the summer months of 1921 and 1922.

From the data accumulated in these cities, comprehensive information has been obtained as to the duration of conditions of average and maximum severity which occur during the humid months. It is of interest to compare the values of the central office conditions of temperature and relative humidity obtained from the recorders, with the corresponding Weather Bureau observations. The curves given in Fig. 6 show a typical comparison from data obtained at Savannah, Ga., during May, 1922. Study of these curves shows that the indoor temperature averaged somewhat higher than that out of doors, and that the indoor relative humidities were seldom higher than 75%, although the outdoor humidities often were higher than 85% for considerable lengths of time. The Weather Bureau data indicate very definitely when rain storms occurred and also periods of high humidity, due perhaps to foggy weather, although such periods are not well defined by the humidity curves showing the indoor conditions.

Since for a given absolute humidity, the relative humidity varies inversely with the change in temperature of the air, obviously it should be possible to keep the relative humidity in a central office building lower than that of the outside air by keeping the windows closed during periods of sudden temperature changes, and by the use of heat in switchboard sections. This latter remedy for humidity troubles has been successfully applied for several years to switchboards installed in some localities. Also the effects upon the indoor humidity and upon the performance of central office equipment, of closing the windows of central office rooms has been the subject of considerable investigation.

In the study of this method of reducing relative humidity, it is very desirable to have records which will show continuously the differences existing between indoor and outdoor temperatures and relative humidities, and in particular to study the effects on the indoor conditions when sudden changes in atmospheric conditions occur such as rain storms when the relative humidity outside reaches 100%. It was found that the automatic recorder described above would lend itself admirably to the study of this problem and that by the use of a simple relay switching mechanism on the recorder, two wind tunnel equipments could be operated with one recorder, enabling temperatures and differences between dry and wet bulb temperatures to be recorded alternately on the same chart for both indoor and outdoor conditions.

A recorder of this type was operated during the summer months of 1921 at the West Street laboratories of the Western Electric Co., Inc., to record the conditions in a well ventilated laboratory room

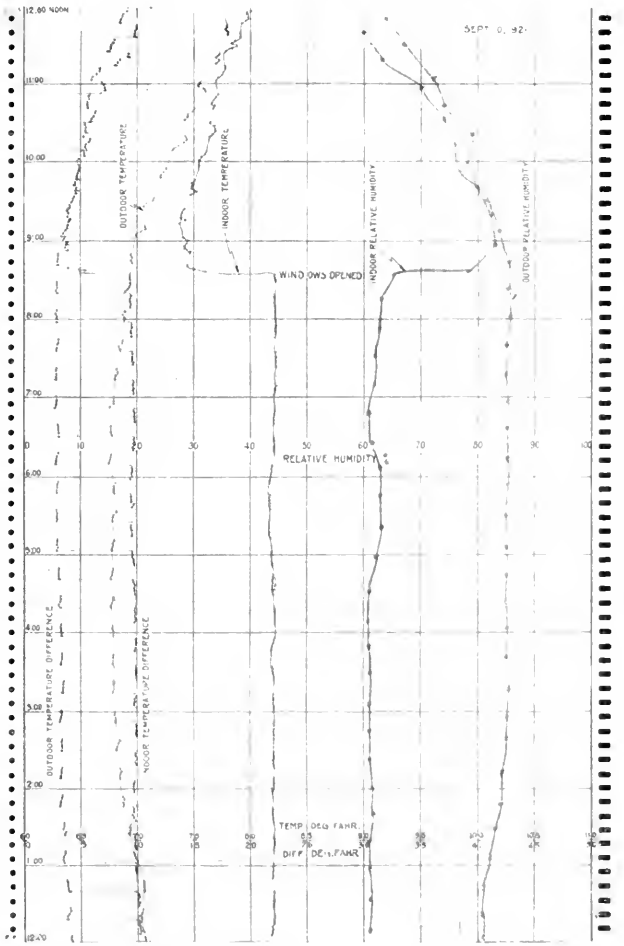


Fig. 7—Record from Double Recorder Comparing Indoor and Outdoor Conditions

about 25 feet x 27 feet and having two windows each in the east and south walls. The wind tunnel equipment was installed at a height of six feet upon a pillar in the center of the room. About ten people normally work in this room. The outdoor conditions were obtained by mounting a wind tunnel equipment in a standard Weather Bureau instrument shelter placed at the top of a tower, 11 feet high, which stands on the roof of a three story building far enough away from walls and other obstacles to permit free circulation of the air.

Figs. 7 and 8 show two typical 12 hour records upon which the indoor and outdoor relative humidities have been plotted from the curves of temperatures and temperature differences recorded by the instrument. A study of these records indicates that large differences often exist between the indoor and outdoor conditions and that the indoor conditions are much less severe than might be expected when the outdoor humidity is high. This difference is particularly noticeable when the windows are closed, but as soon as they are opened the indoor temperature decreases and the humidity generally increases to practically the same value as that of the outside air. Fig. 8 is of particular interest in showing the rapid decrease in the outdoor temperature and increase in relative humidity due to a thunderstorm.

The analysis of the records obtained from a number of recorders which record temperature and difference between dry and wet bulb temperature requires considerable labor in obtaining the corresponding relative humidities from the psychrometric tables and, obviously, periodic values only can be taken unless some rapid mechanical method of doing this is employed. Such methods have been developed and used successfully for this purpose.

A New Direct Reading Humidity Recorder

A much more satisfactory type of recorder is one which, in addition to tracing the temperature curve, traces a curve of the relative humidity. The only instrument of any prominence that has been used in this way is the recording hair hygrometer, the objectionable features of which have already been mentioned.

An improved type of direct reading humidity recorder which has been developed by E. B. Wood, of the Laboratories of the American Telephone and Telegraph Company and the Western Electric Company, employs the Leeds & Northrup automatic recorder mechanism, to which has been added an electrical mechanism which will be described, together with the principle upon which its operation is based.

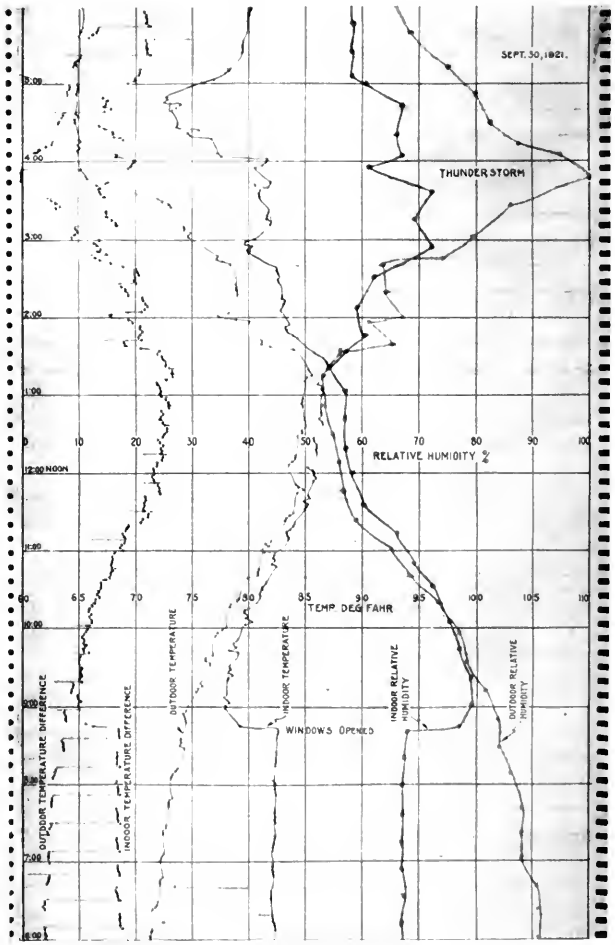


Fig. 8—Record from Double Recorder Comparing Indoor and Outdoor Conditions

This novel improvement depends, for its operation, on the approximate linearity and common intersection of the ordinary humidity curves as shown in Fig. 9.[1]

It is apparent that each of the humidity curves is in effect a straight line and that, with an accuracy sufficient for practical purposes, these curves, representing humidities of from 30% to 100%, converge at a point (a) whose coordinates are (b, c). Assuming that the humidity

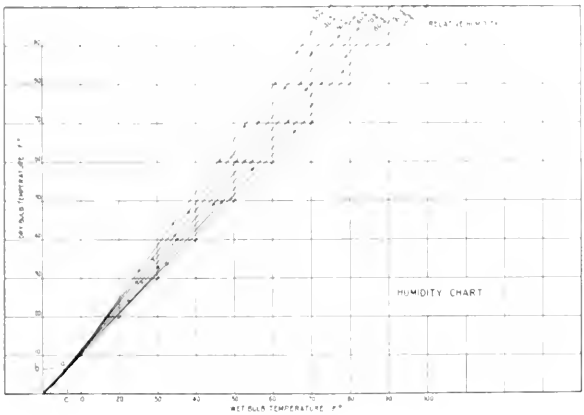


Fig. 9

curves are straight lines passing through point (a), it is apparent that the value of humidity is completely determined if the slope of the particular curve is known, since each curve represents only one value of humidity. It also is apparent that the slope is given by the ratio of dry bulb temperature minus the ordinate of point (a), to wet bulb temperatures minus the abscissa of point (a); or in other words, the relative humidity is completely determined, if the dry bulb and wet bulb temperatures are each known, above the datum coordinates (b, c) of point (a).

If then, a resistance is set off, proportional to the difference between the temperature of the dry bulb and temperature (b), and another resistance is set off proportional to the difference between the temperature of the wet bulb and temperature (c), the ratio between

[1] Bur. Stands. Cir. No. 55, p. 116.

these two resistances will indicate directly the relative humidity corresponding to the dry and wet bulb temperatures. The circuit arrangement by means of which this is accomplished is shown in Fig. 10, and the mechanism of the recorder employing it, is shown in Fig. 11.

The recorder circuit contains three Wheatstone bridges with one battery and galvanometer which are transferred in rotation from

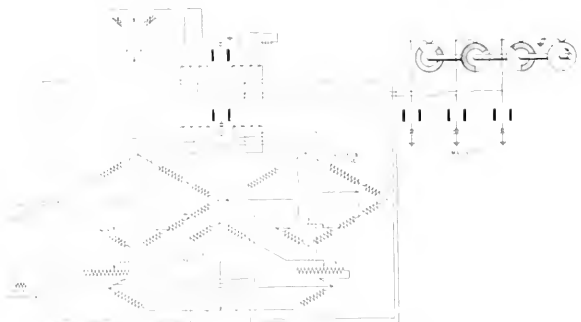


Fig. 10.—Circuit of Direct Reading Recorder

each bridge to the next by the commutator and relays shown in the circuit. The three bridges are arranged so that they remain at their last positions of balance until mechanically connected to the balancing mechanism of the recorder by the electric clutch associated with each bridge whose operation also is controlled by the commutator. The first of these bridges, designated the "dry bulb bridge," contains the dry resistance thermometer and mechanically associated with its slide wire contact is a second slide wire contact operating upon a slide wire resistance arm in the third bridge, designated as the "humidity bridge." The second of these bridges, designated as the "wet bulb bridge," contains the wet resistance thermometer, and mechanically associated with its slide wire contact is a second slide wire contact operating upon a second slide wire resistance arm of the "humidity bridge."

The consecutive balancing of the "dry bulb bridge" and "wet bulb bridge" accordingly sets off resistances upon the two slide wire resistance arms of the "humidity bridge" proportional respectively to the temperature differences described in the second preceding paragraph.

The balancing of this bridge accordingly accomplishes the result already described of determining the ratio of the resistances R_1 and R_2 of these two slide wire arms, and consequently, the relative humidity corresponding to the dry and wet bulb temperatures previously

Fig. 11—Direct Reading Recorder

measured on their corresponding bridges. In the operation of the recorder, a period of about 20 seconds is allowed by the commutator to balance each bridge thus completing a cycle every 60 seconds.

The recorder is equipped with two pens one of which is associated with the slide wire of the "dry bulb bridge" thus recording the dry bulb temperature, while the other pen is associated with the "humidity

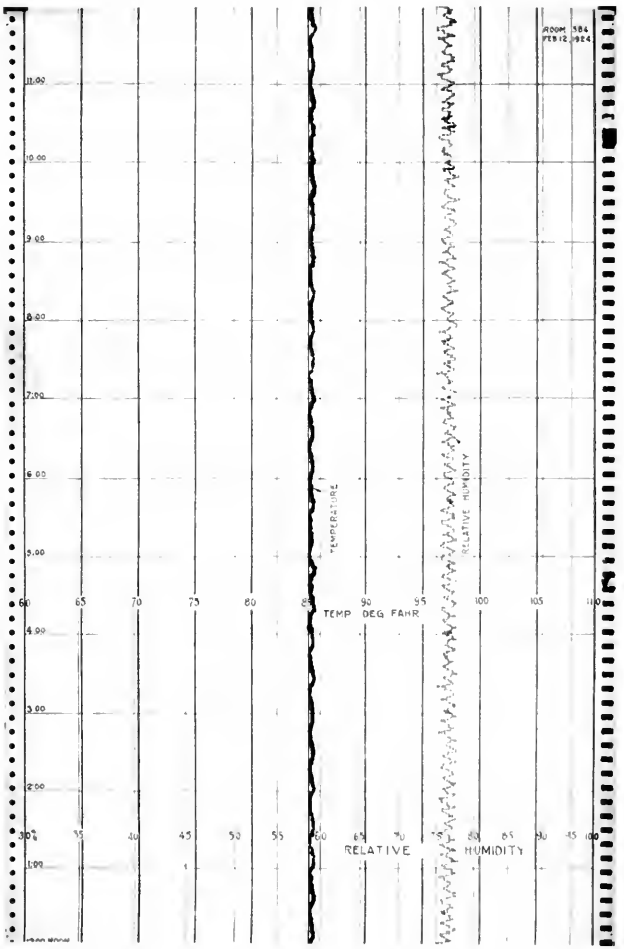


Fig. 12—Temperature and Relative Humidity in a Humidity Room

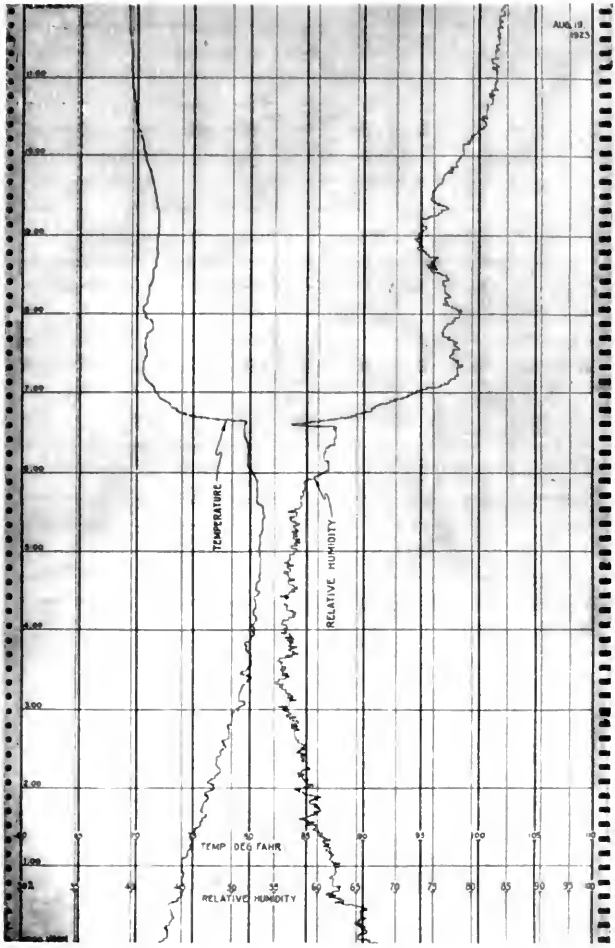


Fig. 13—Outdoor Temperature and Relative Humidity

bridge," thus recording the values of humidity directly. Inasmuch as successive operations of the recorder consist in the restoration of the balance of each bridge, if different from the last position of balance, it is evident that the pens will trace continuously the variations of temperature and relative humidity.

A recorder of this type with its associated wind tunnel mechanism has been used for some time to record the conditions in a laboratory "humidity room." The temperature record given by this recorder is accurate to $\pm 1/4°$ F. as in the case of the difference recorder. The accuracy of the humidity record differs for various points on the scale, depending upon the values chosen for certain resistances in the recorder. When the recorder is adjusted for very close accuracy ($\pm 1/2\%$ relative humidity) for relative humidities above 90%, the accuracy for lower values of humidity decreases until at 50% the maximum variation from the true value may be as much as $2 1/2\%$ relative humidity. If desired, the adjustment may be made to transfer the point of greatest accuracy to any selected lower value of humidity. Experience with this model has suggested changes which should considerably improve this accuracy over the whole range of humidities. Fig. 12 shows a typical 12 hour record of conditions in the "humidity room" while under automatic control of an air conditioning equipment.

This recorder also was used during the summer months of 1923 to record outdoor conditions with the wind tunnel equipment installed in the Weather Bureau instrument shelter mentioned earlier. During this period of 4 months' operation, it required no attention save an occasional oiling of the mechanism and maintenance of the wet bulb equipment, and practically continuous records were secured. The records are of particular interest for observation of the variations of temperature and humidity which take place during changes in weather conditions such as rain storms. Figs. 13 and 14 are reproductions of typical consecutive 12 hour records obtained for outdoor conditions.

From consideration of the humidity recording apparatus which has been developed and the results which have been obtained with it, it may be stated that both the difference recorder and the direct reading recorder are satisfactory instruments with which accurate data may be obtained. However, they are instruments which, in common with other types of apparatus that have been developed to measure humidity, require careful attention of the wind tunnel equipment in order to secure reliable results; also the recorder mechanism itself requires the attention of an operator skilled in its maintenance.

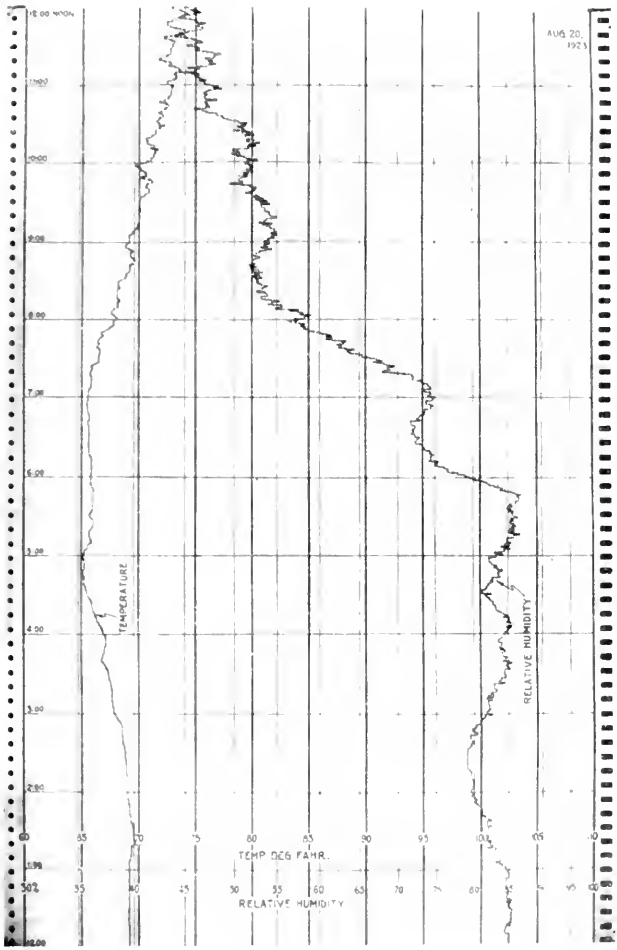


Fig. 14 Outdoor Temperature and Relative Humidity

While the mechanism of the direct reading recorder is more complicated than that of the difference recorder, it is a more useful instrument, both because the humidities may be read directly, thus saving the labor of interpretation of the records, and because the records are more significant. The direct reading recorder, furthermore, may be used to control the functioning of air conditioning apparatus at any desired conditions, at the same time that it is actually recording these conditions. Accordingly, it should prove particularly useful in maintaining proper humidity in apparatus and operating rooms.

A Reactance Theorem

By RONALD M. FOSTER

Synopsis. The theorem gives the most general form of the driving-point impedance of any network composed of a finite number of self-inductances, mutual inductances, and capacities. This impedance is a pure reactance with a number of resonant and anti-resonant frequencies which alternate with each other. Any such impedance may be physically realized (provided resistances can be made negligibly small) by a network consisting of a number of simple resonant circuits (inductance and capacity in series) in parallel or a number of simple anti-resonant circuits (inductance and capacity in parallel) in series. Formulas are given for the design of such networks. The variation of the reactance with frequency for several simple circuits is shown by curves. The proof of the theorem is based upon the solution of the analogous dynamical problem of the small oscillations of a system about a position of equilibrium with no frictional forces acting.

AN important theorem[1] gives the driving-point impedance[2] of any network composed of a finite number of self-inductances, mutual inductances, and capacities; showing that it is a pure reactance with a number of resonant and anti-resonant frequencies which alternate with each other; and also showing how any such impedance may be physically realized by either a simple parallel-series or a simple series-parallel network of inductances and capacities, provided resistances can be made negligibly small. The object of this note is to give a full statement of the theorem, a brief discussion of its physical significance and its applications, and a mathematical proof.

The Theorem

The most general driving-point impedance S obtainable by means of a finite resistanceless network is a pure reactance which is an odd rational function of the frequency p 2π and which is completely determined, except for a constant factor H, by assigning the resonant and anti-resonant frequencies, subject to the condition that they alternate and include both zero and infinity. Any such impedance may be physically

[1] The theorem was first stated, in an equivalent form and without his proof, by George A. Campbell, *Bell System Technical Journal*, November, 1922, pages 23, 26, and 30. By an oversight the theorem on page 26 was made to include unrestricted dissipation. Certain limitations, which are now being investigated, are necessary in the general case of dissipation. The theorem is correct as it stands when there is no dissipation, that is, when all the R's and G's vanish; this is the only case which is considered in the present paper.

A corollary of the theorem is the mutual equivalence of simple resonant components in parallel and simple anti-resonant components in series. This corollary had been previously and independently discovered by Otto J. Zobel as early as 1919, and was subsequently published by him, together with other reactance theorems, *Bell System Technical Journal*, January, 1923, pages 5–9.

[2] The driving-point impedance of a network is the ratio of an impressed electromotive force at a point in a branch of the network to the resulting current at the same point.

constructed either by combining, in parallel, resonant circuits having impedances of the form $iLp+(iCp)^{-1}$, *or by combining, in series, anti-resonant circuits having impedances of the form* $[iCp+(iLp)^{-1}]^{-1}$. *In more precise form,*

$$S=-iH\frac{(p_1^2-p^2)\ (p_3^2-p^2)\ \ldots\ (p_{2n-1}^2-p^2)}{p(p_2^2-p^2)\ \ldots\ (p_{2n-2}^2-p^2)}, \tag{1}$$

where $H\geq 0$ *and* $0=p_0\leq p_1\leq p_2\leq\ \ldots\ \leq p_{2n-1}\leq p_{2n}=\infty$.[3] *The inductances and capacities for the* n *resonant circuits are given by the formula,*

$$L_j=\frac{1}{C_jp_j^2}=\left(\frac{ipS}{p_j^2-p^2}\right)_{p=p_j}\quad (j=1,\ 3,\ \ldots,\ 2n-1), \tag{2}$$

and the inductances and capacities of the $n+1$ *anti-resonant circuits are given by the formula,*

$$C_j=\frac{1}{L_jp_j^2}=\left(\frac{ip}{S(p_j^2-p^2)}\right)_{p=p_j}\quad (j=0,\ 2,\ 4,\ \ldots,\ 2n-2,\ 2n), \tag{3}$$

which includes the limiting values,

$$C_0=\frac{p_2^2\ \ldots\ p_{2n-2}^2}{Hp_1^2p_3^2\ \ldots\ p_{2n-1}^2},\quad L_0=\infty,\ C_{2n}=0,\ L_{2n}=H.$$

Formula (1) may be stated in several mutually equivalent forms.[4] This particular form is the driving-point impedance of the most general symmetrical network in which every branch contains an inductance and a capacity in series, with mutual inductance between each pair of branches. This includes as special cases the driving-point impedances of every other finite resistanceless network.

[3] Since the impedance S is an odd function of the frequency, resonance or anti-resonance for $p=P$ implies resonance or anti-resonance for $p=-P$. In enumerating the resonant and anti-resonant frequencies it is customary, however, to exclude negative values of the frequency. Thus, in the present case, we say that there are n resonant points $(p_1,\ p_3,\ \ldots,\ p_{2n-1})$ and $n+1$ anti-resonant points $(p_0=0,\ p_2,\ p_4,\ \ldots,\ p_{2n-2},\ p_{2n}=\infty)$.

[4] The expression for S given by formula (1) may be written in the mutually equivalent forms,

$$\left[-iH\frac{(p_1^2-p^2)\ (p_3^2-p^2)\ \ldots\ (p_{2n-1}^2-p^2)}{p\ (p_2^2-p^2)\ \ldots\ (p_{2n-2}^2-p^2)}\right]^{\pm 1}\quad\text{and}\quad\left[iHp\frac{(p_2^2-p^2)\ \ldots\ (p_{2n-2}^2-p^2)}{(p_1^2-p^2)\ \ldots\ (p_{2n-3}^2-p^2)}\right]^{\pm 1}$$

If the constant H and all the p_j's of these formulas are restricted to finite values greater than zero, the four cases, obtained by separating the plus and minus exponents, are mutually exclusive, but together they cover the entire field. If p_1 is allowed to be zero, either the first or the second pair covers the entire field. Finally, if in addition p_{2n-1} or p_{2n-2} is allowed to become infinite, while Hp_{2n-1}^2 or Hp_{2n-2}^2 is maintained finite, any one of the four expressions covers the entire field. Sometimes one, sometimes another way of covering the field is the more convenient. Formulas (2) and (3) apply to all of these expressions for S provided the p_j's include all the resonant points and all the anti-resonant points, respectively.

Physical Discussion

The variation of the reactance $X = S\ i$ with frequency is illustrated by the curves of Fig. 1 in all the typical cases of formula (1) for $n = 1$ and for $n = 2$. For every curve the reactance increases with the frequency,[5] except for the discontinuities which carry it back from a positive infinite value to a negative infinite value at the anti-resonant points. Thus between every two resonant frequencies there is an anti-resonant frequency, no matter how close together the two resonant frequencies may be. The effect of increasing n by one unit is to add one resonant point, and thus to introduce one additional branch to the reactance curve, this branch increasing from a negative infinite value through zero to a positive infinite value.

That formula (1) includes several familiar circuits is seen by considering the most general network with one mesh, that is, an inductance and a capacity in series, with the impedance $iLp + (iCp)^{-1}$. This expression is given immediately by (1) upon setting $n = 1$, $H = L$, and $p_1 = 1/\sqrt{LC}$. Since L and C are both positive these constants satisfy the conditions stipulated under (1), thus verifying the theorem for circuits of one mesh. This general one-mesh circuit includes as special cases a single inductance L by setting $H = L$ and $p_1 = 0$, and a single capacity C by setting $H = 0$ and $p_1 = \infty$ such that $Hp_1^2 = 1/C$.

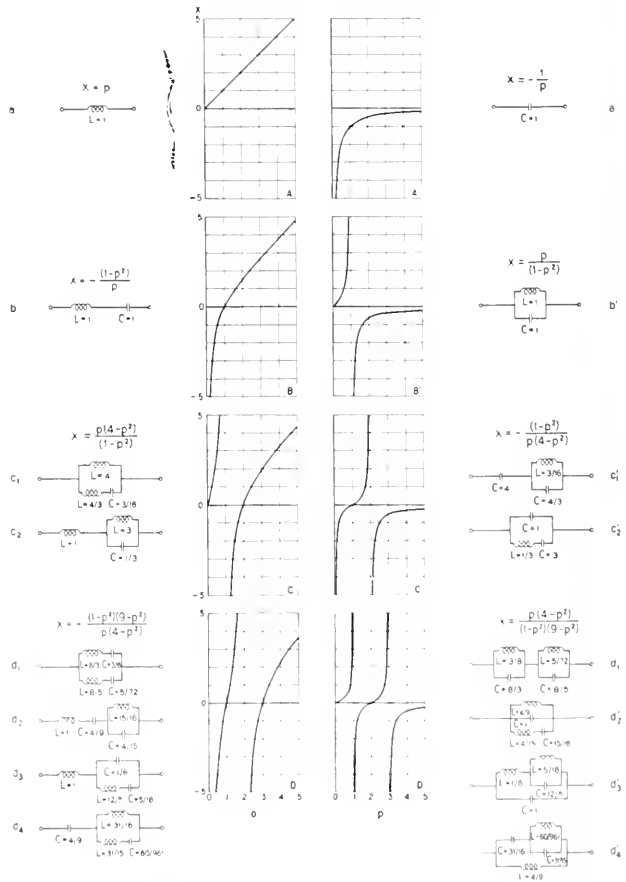
In Fig. 1 the reactances shown by the curves on the right are the negative reciprocals of those on the left. Fig. 1 also shows networks which give the several reactance curves, the networks being computed by means of formulas (2) and (3). The networks are arranged in pairs with reciprocal driving-point impedances and with the networks themselves reciprocally related, that is, the geometrical forms of the networks are conjugate,[6] and inductances correspond to capacities of the same numerical value and vice versa. This relation is a natural consequence of the reciprocal relation between an inductance and a capacity of the same numerical value, these being the elements from which the networks are constructed.

For $n = 1$, formulas (2) and (3) give identical networks, as illustrated by the reactances A, B, A', and B' of Fig. 1, each of which is realized by a single network. For the reactances C and C' the two formulas give distinct networks, c_1 and c_2, c_1' and c_2', respectively, these

[5] This has been proved by Otto J. Zobel (loc. cit., pp. 5, 36), using the formula for the most general driving-point impedance given by George A. Campbell (loc. cit., p. 30).

[6] For a further treatment of conjugate or inverse networks, see P. A. MacMahon, *Electrician*, April 8, 1892, pages 601, 602, and Otto J. Zobel, loc. cit., pages 5, 36, and 37.

two being the only networks with the minimum number of elements which give the specified impedance. In general, however, there are four ways of realizing a given impedance when $n=2$, as illustrated by D and D' of Fig. 1; formulas (2) and (3) give only the first two



1 Reactance curves and networks for simple cases of formula (1).

networks, d_1 and d_2, d_1' and d_2', respectively. The total number of possible ways of realizing a given impedance increases very rapidly for values of n greater than 2; for $n=3$, there are, in general, 32 distinct networks giving a specified impedance.

Formulas (2) and (3) are to be used for determining the constants of the circuits which have certain specified characteristics, whereas most network formulas are for the determination of the characteristics of the circuit from the given constants of the circuit. The application of these formulas is illustrated by the following numerical problem:

To design a reactance network which shall be resonant at frequencies of 1000, 3000, 5000, and 7000 cycles; anti-resonant at 2000, 4000, and 6000 cycles, as well as at zero and infinite frequencies; and have a reactance of 2500 ohms at a frequency of 10,000 cycles.

By formula (1) the reactance of such a network must be

$$X = -H\frac{(p_1^2-p^2)\,(p_3^2-p^2)\,(p_5^2-p^2)\,(p_7^2-p^2)}{p\,(p_2^2-p^2)\,(p_4^2-p^2)\,(p_6^2-p^2)}, \tag{4}$$

where p_1, p_3, p_5, and p_7 are determined by the resonant frequencies to be $1000\times 2\pi$, $3000\times 2\pi$, $5000\times 2\pi$, and $7000\times 2\pi$, respectively; p_2, p_4, and p_6 are determined by the anti-resonant frequencies to be $2000\times 2\pi$, $4000\times 2\pi$, and $6000\times 2\pi$, respectively; and H must be made equal to 0.0596 in order that the reactance at $p=10{,}000\times 2\pi$ may be 2500. The variation of the reactance with the frequency is shown by the curve of Fig. 2.

A network having this reactance may be constructed by combining $n=4$ simple resonant circuits in parallel, or $n+1=5$ simple anti-resonant circuits in series. These two networks are shown by Fig. 2. The numerical values of the elements are determined as follows: Applying formula (2) we have

$$L_1 = \frac{1}{C_1 p_1^2} = H\frac{(p_1^2-p_2^2)\,(p_1^2-p_3^2)\,(p_1^2-p_4^2)}{(p_1^2-p_2^2)\,(p_1^2-p_3^2)\,(p_1^2-p_4^2)} = 0.349,$$

$$L_3 = \frac{1}{C_3 p_3^2} = H\frac{(p_3^2-p_1^2)\,(p_3^2-p_2^2)\,(p_3^2-p_4^2)}{(p_3^2-p_1^2)\,(p_3^2-p_2^2)\,(p_3^2-p_4^2)} = 0.323,$$

$$L_5 = \frac{1}{C_5 p_5^2} = H\frac{(p_5^2-p_1^2)\,(p_5^2-p_3^2)\,(p_5^2-p_4^2)}{(p_5^2-p_1^2)\,(p_5^2-p_3^2)\,(p_5^2-p_4^2)} = 0.264,$$

$$L_7 = \frac{1}{C_7 p_7^2} = H\frac{(p_7^2-p_1^2)\,(p_7^2-p_3^2)\,(p_7^2-p_4^2)}{(p_7^2-p_1^2)\,(p_7^2-p_3^2)\,(p_7^2-p_4^2)} = 0.112;$$

and applying formula (3) we have

$$C_0 = \frac{p_2^2 p_4^2 p_6^2}{H p_1^2 p_3^2 p_5^2 p_7^2} = 0.0888 \times 10^{-6}, \; L_0 = \infty,$$

$$C_2 = \frac{1}{L_2 p_2^2} = \frac{-p_2^2(p_4^2 - p_2^2)(p_6^2 - p_2^2)}{H(p_1^2 - p_2^2)(p_3^2 - p_2^2)(p_5^2 - p_2^2)(p_7^2 - p_2^2)} = 0.0461 \times 10^{-6},$$

$$C_4 = \frac{1}{L_4 p_4^2} = \frac{-p_4^2(p_2^2 - p_4^2)(p_6^2 - p_4^2)}{H(p_1^2 - p_4^2)(p_3^2 - p_4^2)(p_5^2 - p_4^2)(p_7^2 - p_4^2)} = 0.0523 \times 10^{-6},$$

$$C_6 = \frac{1}{L_6 p_6^2} = \frac{-p_6^2(p_2^2 - p_6^2)(p_4^2 - p_6^2)}{H(p_1^2 - p_6^2)(p_3^2 - p_6^2)(p_5^2 - p_6^2)(p_7^2 - p_6^2)} = 0.0725 \times 10^{-6},$$

$$C_8 = 0, \qquad L_8 = H = 0.0596.$$

These formulas give the numerical values of the inductances in henries and the capacities in farads. The entire set of numerical values is shown in Fig. 2. It is to be noted that the anti-resonant circuit corresponding to $p_0 = 0$ consists of a simple capacity since the inductance is infinite and thus does not appear in the network, whereas for $p_8 = \infty$ the anti-resonant circuit consists of a simple inductance, the capacity being zero and thus not appearing in the network.

Mathematical Proof

We shall first prove that the driving-point impedance S, as given by (1), may be physically realized by either a simple parallel-series or a simple series-parallel network of inductances and capacities, provided resistances can be made negligibly small.

The rational function $1/S$ can be expanded in partial fractions,

$$\frac{1}{S} = \frac{iH_1 p}{p_1^2 - p^2} + \frac{iH_3 p}{p_3^2 - p^2} + \cdots + \frac{iH_{2n-1} p}{p_{2n-1}^2 - p^2},$$

where

$$H_j = \left(\frac{p_j^2 - p^2}{ipS} \right)_{p = p_j} \quad (j = 1, 3, \ldots, 2n-1).$$

Hence S is equal to the impedance of the parallel combination of the n circuits having the impedances $(p_j^2 - p^2)/(iH_j p) = iH_j^{-1} p + [i(H_j p_j^2)^{-1}] p]^{-1}$, that is, n simple resonant circuits in parallel, each circuit consisting of an inductance and a capacity in series, with the numerical values given by (2). Furthermore, these numerical values of the inductances and capacities given by (2) are all positive, an even number of negative factors being obtained upon substituting $p = p_j$, since in every case $p_j \leq p_{j+1}$. Hence the network defined by (2) has the impedance S as given by (1) and is physically realizable.

Likewise, by expanding S in partial fractions, it can be shown that the network defined by (3) has the impedance S as given by (1) and is physically realizable.

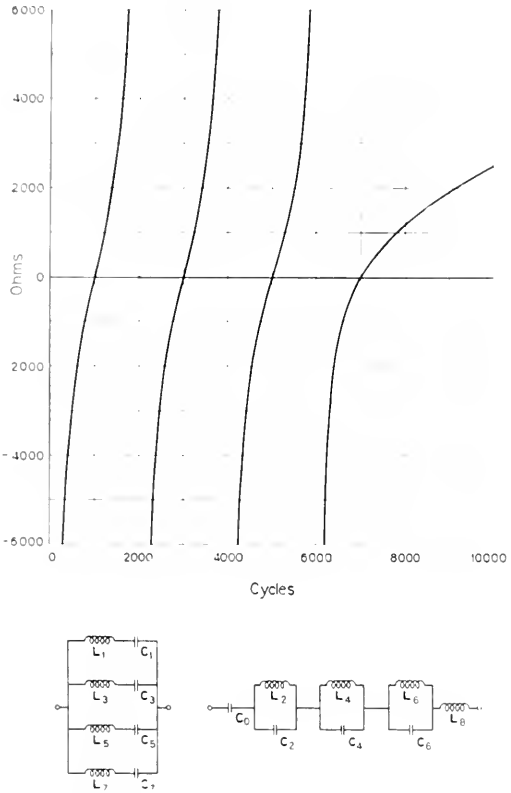


Fig. 2.—Reactance curve and networks for formula 4.

The values of the inductances and capacities are in henries and microfarads

$L_1 = 0.349$	$C_1 = 0.0726$	$L_2 = 0.137$	$C_0 = 0.0888$
$L_3 = 0.323$	$C_3 = 0.00872$	$L_4 = 0.0302$	$C_2 = 0.0461$
$L_5 = 0.264$	$C_5 = 0.00384$	$L_6 = 0.00971$	$C_4 = 0.0523$
$L_7 = 0.142$	$C_7 = 0.00363$	$L_8 = 0.0596$	$C_6 = 0.0725$

The electrical problem of the free oscillations of a resistanceless network is formally the same as the dynamical problem of the small oscillations of a system about a position of equilibrium with no frictional forces acting. The proof of formula (1) may be derived from the treatment of this dynamical problem as given, for example, by Routh.[7]

In any network the driving-point impedance in the qth mesh, S_q, is equal to the ratio A/A_q, where A is the determinant[8] of the network and A_q the principal minor of this determinant obtained by striking out the qth row and the qth column. The determinant of a network has the element Z_{jk} in the jth row and kth column, Z_{jk} being the mutual impedance between meshes j and k (self-impedance when $j=k$), the determinant including n independent meshes of the network.

Hence the determinant A has the element $Z_{jk}=iL_{jk}p+(iC_{jk}p)^{-1}$, where L_{jk} is the total inductance and C_{jk} the total capacity common to the meshes j and k. Upon taking the factor $(ip)^{-1}$ from each row and substituting $-p^2=x$, the expression for A may be put in the form $A=(ip)^{-n}D$, where D is a determinant with $L_{jk}x+1/C_{jk}$ as the element in the jth row and the kth column. This is of exactly the same form as the determinant given by Routh[9] for the solution of the dynamical problem; it is proved there that this determinant, regarded as a polynomial, has n negative real roots which are separated by the $n-1$ negative real roots of every first principal minor of the determinant.

Hence, we may write $D=E(x_1+x)(x_3+x)\ldots(x_{2n-1}+x)$, where $x_1, x_3, \ldots, x_{2n-1}$ are all positive and arranged in increasing order of magnitude, and where E is also positive since D must be positive for $x=0$. The determinant D_q may be expressed in similar manner since it is of the same form as D but of lower order.

[7] E. J. Routh, "Advanced Rigid Dynamics," sixth edition, 1905, pages 44–55. In the notation of the dynamical problem as presented here, the coefficients A_{jk} correspond to the inductances, $1/C_{jk}$ to the capacities, $p/(2\pi)$ to the frequency, and θ', ϕ', etc., to the branch currents in the electrical problem.

A complete proof of formula (1) has been worked out for the electrical problem, without depending in any way upon the solution of the corresponding dynamical problem. This proof has not been published here in view of the great simplification made by using the results already worked out for the dynamical problem.

[8] A complete discussion of the solution of networks by means of determinants has been given by G. A. Campbell, Transactions of the A. I. E. E., 30, 1911, pages 873–909.

[9] The determinant given by Routh (loc. cit., p. 49) has the element $A_{jk}p^2+C_{jk}$.

The driving-point impedance is given by

$$S_z = \frac{A}{A_q} = (ip)^{-1}\frac{D}{D_q} = (ip)^{-1}\frac{E(x_1+x)(x_3+x)\cdots(x_{2n-1}+x)}{E_q(x_2+x)\cdots(x_{2n-2}+x)},$$

where $0 \leq x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_{2n-2} \leq x_{2n-1}$, since the roots of D are separated by the roots of D_q. Upon substituting $x = -p^2$ and introducing the notation $H = E/E_q$ and $p_1^2, p_2^2, \ldots, p_{2n-1}^2 = x_1, x_2, \ldots, x_{2n-1}$, respectively, we see that formula (1) is completely verified as the most general driving-point impedance obtainable by means of a finite resistanceless network.

Some Contemporary Advances in Physics—III

By KARL K. DARROW

ELECTROMAGNETIC waves of every frequency from 10^4 to 10^{20} exist; they can be generated and perceived; their frequencies in nearly every instance can be measured; their actions and reactions with matter can be studied. This brief statement is the synthesis of a great multitude of inventions, experiments and observations upon phenomena of extraordinary diversity and variety. When Herschel in 1800 carried a thermometer across the fan-shaped beam of colored light into which a sunbeam was resolved by a prism, and observed that the effect of the sunbeam on the mercury column did not cease when it passed beyond the red edge of the fan, he proved that the boundary of the spectrum beyond the red is imposed by the limitations of the eye and not by a deficiency of rays. Almost at the same time Ritter found that the power of the violet rays to affect salts of silver was shared by invisible rays beyond the violet edge of the beam. Maxwell developed the notion of electromagnetic waves from his theory of electricity and magnetism, and described some of the properties they should have; and the light-waves and the infra-red and ultra-violet rays were found to have some of these properties, while the outstanding discordances were explained away by Maxwell's successors. Hertz and many others built apparatus for producing Maxwell's waves with frequencies far below those of light, and apparatus for detecting them, with consequences known to everyone. Years after X-rays and gamma-rays were discovered emanating from discharge-tubes and disintegrating atoms, Laue proved that these too are waves, lying beyond the visible spectrum in the range of high frequencies. Radiations emerging from collapsing atoms and radiations diverging from wireless towers; waves conveying the solar heat and waves carrying the voice; rays which disrupt atoms by extracting their electrons, rays which alter atoms by rearranging their electrons, rays which almost ignore atoms altogether, were successively discovered or created; and all these radiations were brought into one class, and identified with light.

This enormously extended electromagnetic spectrum was interrupted until lately by two regions unexplored. They were known as the gap between the X-rays and the ultra-violet, and the gap between the infra-red and the Hertzian waves, according to the names by which the various explored regions of the spectrum commonly go; but to understand why they remained unclosed for so long, and what kinds of rays are being found within them, it is necessary to consider

how certain properties of the waves vary along the spectrum. Enough is known about the origin of electromagnetic waves to justify using it as a basis of classification. Classifying the rays, therefore, by *mode of production*, we can distinguish at least four sharply-contrasted types: first, rays emitted from atomic nuclei in process of disintegration; second, rays emitted from atomic electron-systems in process of rearrangement; third, rays due to atoms vibrating to and fro about their positions of equilibrium as constituents of molecular groups or of space-lattices; and finally, waves generated by oscillating electrical circuits.[1] For each of these classes there is a region of the spectrum which is particularly, although not exclusively, its own.

The rays emitted from disintegrating nuclei lie at the topmost end of the frequency-scale; they overlap the rays of the second class, but do not approach either of the gaps. The rays resulting from rearrangements of the electron-systems surrounding atom-nuclei extend over an enormous range. The minimum wave-length of this range is .1075A, the *K*-frequency of the uranium atom; it is and will almost certainly remain the definitive limit, unless someone should succeed in discovering a substance further up the periodic table than uranium, or in removing some of the deepest electrons from the electron-system of some heavy[2] atom. As maximum wave-length we might take that of a line 40500A lately recognized by Brackett as belonging to atomic hydrogen; but this is certainly not the definitive limit. Emission-bands due to atoms vibrating within molecular groups are found in and beyond the "near infra-red" (and indeed in the ultra-violet around 3000A, if we include bands of "compound" origin, resulting from processes occurring together which if happening separately would produce rays of the second and third types, respectively); while the "residual rays," which are ascribed to atoms vibrating within the gigantic molecular group which is a crystal lattice, extend as far as 0.152 mm. (residual rays of thallium iodide). Between 0.1 mm. and 0.4 mm. rays have been discovered emanating from the mercury

[1] This classification is obviously not an exhaustive one. Continuous spectra have been omitted—thermal emission spectra of solids, and continuous X-ray spectra, which may be ascribed to random accelerations of free electrons. The continuous bands in gas spectra, of which one has just been explained by Gerlach (*ZS. f. Phys.*, 18, pp. 239–248; 1923, and others by Bohr (*Phil. Mag.* 26, p. 17; 1913), can be included in the second class by a slight generalization; and so, probably, can some fluorescence and phosphorescence spectra, at least if we extend "atomic electron-systems" to include "electron-systems of grouped atoms." There is also the possibility of rays due to changes in rate of rotation of molecules, not compounded with changes in oscillation or electron-arrangement.

[2] Meaning an atom with a large nuclear charge, which would have heaviness, or more properly massiveness, as a secondary characteristic. A short and simple adjective to describe where an atom stands in the scale of nuclear charge, i.e. in the periodic table, would be very welcome.

arc, which probably belong to the second or third class, but it is not certain which. If we gather all these classes together into a single great class of *natural* rays, extending from .02A or 2.10^{-10} cm. to 4,000,000A or 0.04 cm., they may be contrasted with the *artificial* rays generated by man-made electrical circuits, lying entirely beyond the long-wave limit of their range.[3]

One of the two lacunae in the spectrum, extending from 0.4 mm. to 7 mm., separated the range of natural rays from the range of artificial rays. To close this gap it was necessary literally to invent new rays, by designing oscillating electrical circuits which would generate frequencies which perhaps had never existed before in nature. The other lacuna, extending from 13A to 1200A, lay by contrast in the very centre of the range of natural rays, and precisely where we expect to find the frequencies resulting from certain peculiarly interesting and important processes in the electron-systems of atoms. These processes, it appears, are not in all cases easy to incite by the usual methods of stimulating atoms to radiate; but this difficulty is only one, and probably the least serious one, of the three hindrances which combined to delay the exploration of this region. A second impediment comes from the limitations of our devices for measuring wave-length, every one of which is unavailable over a certain sector of the region, extending roughly from 13A to 150A (limits which may later be forced somewhat closer together); but the most conspicuous obstacle is the extraordinary obstructiveness and opacity of every kind of matter to these rays.

The ability of electromagnetic waves to penetrate matter varies enormously from one part of the spectrum to another. At the uppermost end of the frequency-scale, the rays penetrate every sort of matter with astonishing ease. A layer of lead 8 mm. thick is required to remove half of the energy of a ray of wavelength .025A; and even this, it is probable, is not absorbed in the strict sense of being converted from radiant energy into another form, being merely deflected or *scattered* out of its original direction of motion.[4] With rays of greater wave-length, a true absorption is superposed upon the scattering, and increases very rapidly, about as the third power of the wavelength. The absorbed energy is used in extracting electrons from

[3] The distinction between natural and artificial rays is striking, but I fear not quite exact, since lightning-discharges and the causes of "static" offer instances of natural sources of radio frequencies. Also the selective absorptions of certain substances in the Hertzian range strongly suggest natural emission-frequencies. Still the distinction is not yet unsound enough to be dangerous.

[4] If A. H. Compton's theory of X-ray scattering is eventually triumphant, it will be necessary to admit that some radiant energy is transformed into kinetic energy of moving masses when scattering occurs.

the deeper levels of atomic electron-systems, as I described in the second of these articles. The absorption in any particular substance does not increase with an uninterrupted upward sweep; there are occasional setbacks, each of which occurs at a critical frequency where the radiation ceases to be able to extract electrons from a particular level. But though the lower-frequency rays cannot extract the deeper electrons of the atoms, they more than make up for it by expelling the outer electrons in greater and greater abundance; and when wave-length 13A is reached, they can remove only the outermost electrons or shift them from one orbit to another,[5] but they perform these actions so often that the beam is rapidly absorbed (even at 0.5A, 0.01 mm. of lead is sufficient to abstract half its energy).

Beyond 13A there is a region of well-nigh total eclipse. All we know about it is derived from a few measurements by Holweck. According to him, rays of wave-length 10A lose half their energy in traversing half a millimetre of air at atmospheric density; at 100A, the same proportion is consumed in a twentieth of a millimetre of air, or in a quarter of a millimetre of hydrogen, the most tenuous of all substances; and even these are not the most absorbable rays. A sheet of celluloid, .0001 mm. thick, which absorbs only 8% of the energy of a beam of wave-length 10A and 36% at 100A, abstracts 91% of the energy at 250A. It actually absorbs 97.3% of a ray of wave-length 308A; but this may be the least penetrating radiation of the entire scale, for the transmission apparently is a little greater at 400A (although Holweck seems to distrust the reliability of the last result). It must be admitted that the various beams of radiation on which these measurements were made are not monochromatic, but comprise each a continuous range of wave-lengths extending down to the quoted value, which is the minimum. Since the beam is in every case filtered through as many absorbing layers as possible before the final measurement of transmission through the celluloid sheet is made, and those remove preferentially the longer waves, it is probable that each datum refers to a finite, yet comparatively narrow, band of wave-lengths with its lower end at the specified value.[6]

[5] Some of the absorbed energy may be utilized in other ways, but there is no known alternative mechanism.

[6] The curve of Fig. 1, taken from Holweck's article, shows his data for the absorbing power of celluloid plotted logarithmically against wave-length. All the points refer to wave-lengths between 40A and 400A except the one marked "a," which refers to the rays emitted by gaseous hydrogen bombarded by electrons of energy between 13 and 38 volts (the transmission is the same for every bombarding-voltage within this range). It is probably a sort of "weighted-mean" value for the various radiations of the Lyman series and possibly the secondary spectrum of hydrogen, and the value 1140A which Holweck assigns as its effective wave-length is probably as good as any. The straight line on the left relates to nitrogen.

Whether or not Holweck's measurements are accurate enough to fix the point of greatest opacity, it is certain that somewhere between 300A and 1200A the eclipse begins to pass off. Fluorite commences to transmit at about 1200A, quartz and gelatine at about 1800A (each

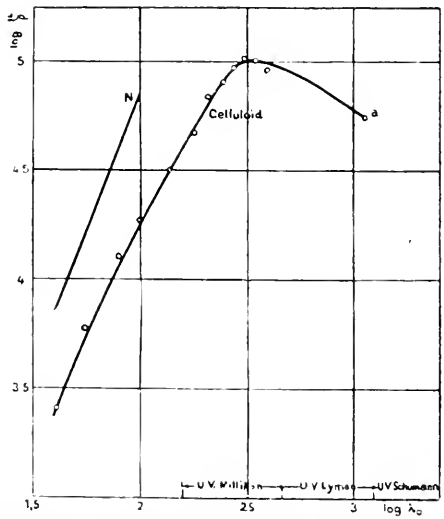


Fig. 1—Absorbing power of celluloid and of nitrogen plotted versus wave-length in the region of greatest opacity. (*Annales de Physique.*)

of these was, for reasons of experimental technique, long the limit of the explored region). Air begins to let through the light at about 1800A; the atmosphere indeed arrests the rays of the sun and stars as far along as 2900, but this is ascribed to ozone in the upper strata. Henceforward the absorption of radiant energy in gases consists mainly in shifting the valence-electrons of the atoms from one level to another, or in altering the amplitude of vibration of atoms built into molecular groups. The characteristics of individual atoms become steadily less influential; the groupings of the atoms into molecules, crystals, liquid or solid continua determine the amount of absorption. The question whether a particular solid is a conductor or an insulator, entirely irrelevant at high frequencies, eventually becomes the only

question that matters; and radio-frequencies penetrate great thicknesses of rock or brick more readily than the thinnest sheet of metal foil.

To explore the region of the spectrum in which the absorbing-power of matter is at its greatest, it is necessary to make a high vacuum over the entire path of the rays from their source to the receiver (photographic plate, ionization-chamber, or electrode for photoelectric emission). This necessity can be escaped only if the obligation of measuring wave-lengths is evaded, for then the path may be very short; the receiver may be brought quite close to the piece of solid substance or the stratum of gas in which the rays are excited. If the wave-lengths are measured, it must be done with a ruled or crystalline diffraction-grating, which enforces a lengthy path (often as much as two metres). No solid windows can be interposed in it to confine a diffusing gas to the region where the rays are excited (the only exceptions yet developed are Holweck's .0001-mm. celluloid windows, which when stretched over and sustained by a fine-meshed gauze are said to be able to support a 5-cm. pressure-difference between their two faces). The excitation must therefore take place, whenever possible, in vacuo. This is simple enough when dealing with the rays excited from solids by electron-bombardment, and originating from displacements of electrons deeper down in the atomic system than the valence-electron; for the bombardment can be carried on in vacuo. But the arcs and sparks which are commonly used to displace the valence-electrons of free atoms or molecules, and so produce the frequencies for which these are responsible, are usually operated in an atmosphere composed of a comparatively few of the atoms being studied, mingled with a large amount of air or some other permanent gas. Yet it has been found possible to operate both arc and spark discharges "in vacuo,"-that is, without the atmosphere of permanent gas; though they differ in various ways from the like-named and familiar discharges in air, and do not display quite the same spectra.

Vacuum arcs, when once ignited, can be maintained with a moderate voltage between electrodes of various metals; the mercury vapor lamp is the familiar example, but arcs of such metals as magnesium, aluminium, and lead were developed as early as 1905. The name "vacuum arc" is, of course, a misnomer; the discharge occurs in an atmosphere of the vapor of the metal, but this congeals as soon as it starts to diffuse away from the discharge, and does not impair the vacuum in the light-path. The condition for an easily-maintained vacuum arc is that the vapor-pressure of the metal involved be comparatively high. Yet arcs between carbon electrodes in vacuo seem

to be easy to maintain, though the vapor pressure of carbon is immeasurably small; one is led to suspect the gases inevitably occluded in this element.[7] Saunders produced waves as short as 978A with an arc in calcium vapor, and Simeon waves down to 375A with a "carbon vacuum arc."

These vacuum arcs are started either by heating the electrodes to produce a momentary high vapor-density, and applying a transient high voltage between them; or by touching them together and drawing them apart while the moderate voltage is applied. If the latter method is tried when the voltage is too low to maintain an arc, there is a transitory flash, the *breakspark*; its spectrum in the visible region has been noticed by von Welsbach, who finds the relative intensities of certain lines strangely altered from what they are in the ordinary spark; but according to McLennan and Lang, it yields no rays of wave-length inferior to 2000A.

The *vacuum spark* or *hot spark* employed by Millikan and his associates is an altogether different affair; it is a brilliant spark which occurs between electrodes a millimetre or so apart (the limits 0.1 mm. and 2 mm. have been assigned) in an extremely high vacuum, when a transient potential-difference of the order of several hundreds of thousands of volts is laid across them. This is a mysterious phenomenon, which has been studied by several scientists, without satisfactory conclusions. Whatever the vacuum spark really is, there is no doubt that it exists, and that wave-lengths are found in its spectrum which are shorter than any hitherto observed in any spectrum of arc or spark; and it is likely that these high-frequency rays are not excited at all in the ordinary electrical discharges of relatively low voltage, so that the high vacuum provides the conditions for stimulating as well as for transmitting them. The least wave-length yet measured with an optical method (ruled grating), which is 136A, occurs in the spectra of some of these sparks.

Most difficult of all is obviously the problem of detecting the rays emitted by the atoms or molecules of a permanent gas, which must of necessity occupy the entire path of the light from the place where it is excited to the place where it is received, unless intercepted by a solid partition which would intercept the desired waves also. If the discharge-tube containing the luminous gas communicates only by a narrow slit with the chamber containing the diffracting and receiving apparatus, it is practicable to connect a powerful pump to

[7] The minimum maintaining voltage for arcs in vacuo is given by Simeon as follows, for electrodes of the following materials: *C* 30 to 40 volts, *Na* 30 to 40, *Al* 80 to 100, *Si* 95 to 105. The distance between the electrodes is described as "slight," the degree of vacuum before arcing is not stated.

a branch-tube opening near the slit into the latter chamber, and so maintain in it a considerably lower density of gas than is required in the discharge. Hopfield has succeeded in maintaining an atmosphere of one kind of gas in the discharge-tube, and an atmosphere of another and a more transparent kind of gas in the chamber; the two gases are prevented from mingling by the same pumping-arrangement.

As for the measurement of wave-lengths from 1200Å down to about 100Å, it must be made with a concave diffraction-grating, which separates rays of different wave-lengths and itself focusses them at different places; for the rays cannot penetrate the prism of a prism spectograph, or the lens which is commonly used[8] to focus the beams diffracted by a plane grating. Rowland of Johns Hopkins, the first great master of the art of making diffraction-gratings, ruled them both upon plane and upon concave surfaces. The plane grating was so much the more easily ruled, that the concave grating fell into desuetude; but it became invaluable as soon as Lyman began to work in the region where the lenses extinguish the light. One might have anticipated that it would refuse to diffract rays the wave-lengths of which are only one-twentieth, one-fiftieth, even one one-hundredth of the spacing between its lines; but as Lyman and Millikan advanced farther and farther beyond the earlier limit of the ultra-violet, the concave grating proved itself competent to an extent which would probably have astonished its inventor. In one of Millikan's articles we may read an account of the ruling of new gratings by Pearson of Chicago; the spacing of the lines was by no means unusually small (about 500 per mm.) but they were ruled "with a very light touch so as to leave a portion of the original surface functioning in the production of spectra"—partly so that successive rulings might be nearly alike, but chiefly because if just half the original surface could be left intact, a large proportion of the total radiant energy would be diffracted into the first-order spectrum (this is the only usable one, because the higher-order images formed by the small wave-length rays encroach on the first-order images of the rays of greater wave-lengths). The arrangement of apparatus in experiments with the concave grating has varied little from the form which Lyman originally gave it. In Fig. 2 (from an article by McLennan) one sees the cross-section of a large tubular air-tight chamber, containing the grating at L (it is mounted on a carriage Q sliding on rails O, P), the slit at S and the photographic

[8] There is no apparent reason against using concave mirrors instead of lenses, unless the multiple reflections consume too much of the light. Luckiesh mentions an instrument designed with focussing mirrors of nickel Houston, Proc. Roy. Soc. Edinb., 1912), which, however, were found inferior to quartz lenses in the range in which it was tested.

plate at C. The rays are excited at the centre of a tube V communicating by the slit with the grating-chamber. In this instance the source of light was a vacuum-spark between the electrodes sketched; had it been a vacuum arc or a glow-discharge in a permanent gas, the tube might have been different in appearance, but would have been sealed onto

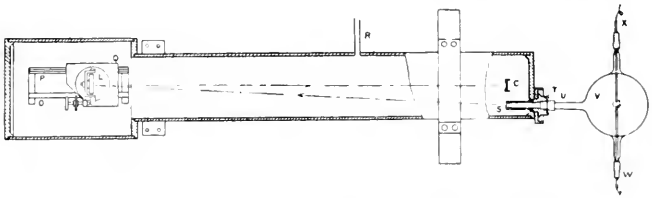


Fig. 2—Vacuum spectrograph with concave grating. (*Proceedings of the Royal Society.*)

the chamber at the slit in the same manner. The distance SL and LC are each one metre, and the sum of them constitutes the major part of the light-path (Lyman has reduced the sum to 40 cm. by using a more curved grating).

The extension of the explored or explorable region of the spectrum from 1200A onward to 136A does not entirely close the lacuna; but it brings into the accessible range every one of a certain very important class of rays—the rays emitted by a free atom when its valence-electron has been displaced and is returning towards or to its normal position. The reason for distinguishing one electron of the atomic electron-system above the others as the *valence electron* (the name is chosen rather for its meaninglessness than for its meaning) lies in the existence of line-series in the spectra. Magnificently regular series of rays are observed in the spectra of the atoms of hydrogen and of ionized helium, each of which has an electron-system consisting of a single electron in the inverse-square field surrounding the atom-nucleus.[9] Series which resemble these, though they are not arranged according to so elegantly simple a numerical law, are found in the spectra of the elements of the first column of the periodic table (Fig. 3) and suggest forcibly that one of the electrons of the atom of lithium, or sodium, or potassium lies so much farther out than all the others that it moves by itself in a field which is almost identical with the inverse-square field of a nucleus of charge e (the resultant of the fields of the nucleus and the inner electrons approaches such a

[9] This inverse-square field seems to be assured by the experiments on deflections of alpha and beta particles by atom-nuclei, quite apart from the successes of Bohr's special assumptions about atomic structure and radiation.

field as the distance from them increases). The same argument applies to elements of the second, third, and fourth columns, though with diminishing force, for the series become more difficult to trace and depart greatly from the archetype. In the crowded and complicated spectra of elements such as neon, argon, and iron, it is very

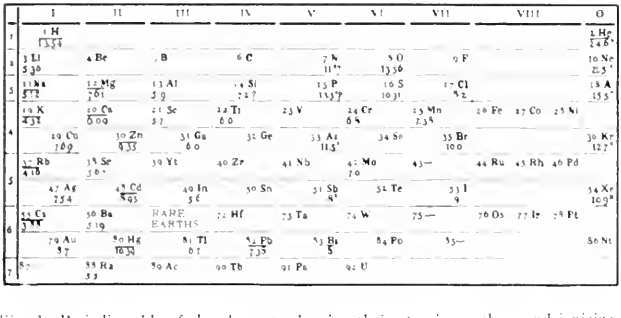


Fig. 3 Periodic table of the elements showing their atomic numbers and ionizing-potentials. Cf. footnote 15.

difficult, though apparently not impossible, to arrange frequencies into series, and this is in accord with the belief (founded on evidence of other kinds) that in these atoms there is no single outer electron far beyond all the others, but rather an outer shell of several similarly-placed electrons. Any one of these might imitate the behavior of a valence-electron, however, when removed to an unusually large distance from the nucleus and from the rest. It is to be observed also that when atoms are brought close together in the liquid or solid state, the line series can no longer be excited.

Wherever, therefore, there are discernible line-series, one infers an electron far enough beyond all the others to have a behavior and deserve a title of its own. Generalizing Bohr's wonderfully successful model of the atoms of hydrogen and ionized helium, we imagine that this electron enjoys a particular set of orbits, in the narrowest and deepest-lying of which it normally abides, while in any one of the others it can make only a transient halt.[10]

[10] It may not be superfluous to complete the description of Bohr's model by saying that when the electron goes from one orbit to another, the difference ΔU between the values of the energy of the atom in the two states is radiated in a ray of frequency $\Delta U / h$.

Now all the line-series observed in the spectra of excited atoms and all which there is any reason to imagine as existent but undiscovered, lie entirely at wave-lengths greater than 136A; indeed most of them lie in the already-accessible region beyond 1200A, but a few of the most important are in the newly-opened range. Hydrogen is entitled to first mention, being the leader of the procession of elements as well as the most completely understood of them. The visible spectrum of (atomic) hydrogen consists of the archetype of all line-series, the Balmer series, extending from 6563A to 3650A, the frequencies of its lines being equal to the numbers of the series

$$(A) \qquad R\left(\frac{1}{2^2}-\frac{1}{3^2}\right),\ R\left(\frac{1}{2^2}-\frac{1}{4^2}\right),\ R\left(\frac{1}{2^2}-\frac{1}{5^2}\right),$$

and so forth, in which R is a certain constant ($R=3.29\cdot 10^{15}$). According to Bohr's theory, this means that the energy-values[11] of the consecutive orbits of the valence-electron (in this case the only electron) are given by the numbers of the succession

$$(B) \qquad -Rh\left(\frac{1}{2^2}\right),\ -Rh\left(\frac{1}{3^2}\right),\ -Rh\left(\frac{1}{4^2}\right),\ -Rh\left(\frac{1}{5^2}\right),$$

and so forth, and the consecutive rays of the series are emitted when the electron drops into the first of these orbits from the second, third, fourth and consecutive orbits. Most people, on looking at the succession of numbers (B), would instinctively complete it by adding a term $-Rh$ at the beginning; and if there is truly an orbit of which the energy-value is $-Rh$ there must be an additional line-series,[12] the frequencies of its lines being equal to the numbers of the series

$$(C) \qquad R\left(1-\frac{1}{2^2}\right),\ R\left(1-\frac{1}{3^2}\right),\ R\left(1-\frac{1}{4^2}\right),\ \text{and so forth.}$$

The first three lines of this series should lie at 1216A, 1026A and 972A. They were discovered by Lyman in 1913, and the series bears his name.

[11] The energy-value of an orbit is the energy of the atom when the valence-electron is in this orbit; the energy of the atom being set equal to zero, when the valence-electron is removed to infinity. It follows from this last convention that the energy-value of an orbit, with sign reversed, is equal to the energy which must be imparted to the atom to remove the valence-electron completely from the atom when it is initially in the orbit in question. Thus the energy-value of the orbit which the valence-electron normally inhabits is equal to the ionizing-potential of the atom, when it is expressed in appropriate units and its sign reversed. The practical advantages of this convention are so great that we endure its annoying and confusing consequence of making all the energy-values of non-ionized atoms negative.

[12] The existence of this series was anticipated long before Bohr's interpretation of the Balmer series, being suggested by the form of the series itself.

Helium follows hydrogen in the procession of elements. Its spectrum includes several line-series. The frequencies of the first four members of one of these series, the principal series of the singlet or parhelium spectrum, are as follows (all the numbers in the successions D, E, F, G, and H should be multiplied by 10^{14}):

(D) 1.457, 5.981, 7.567, 8.300

Subtracting each from the frequency of the series-limit, which is 9.609, we obtain the succession of numbers

(E) $(9.609-8.152)$, $(9.609-3.628)$, $(9.609-2.042)$, $(9.609-1.309)$

which suggests a succession of orbits, having the following consecutive energy values[13]:

(F) $-9.609h$, $-8.152h$, $-3.628h$, $-2.042h$, $-1.309h$.

The consecutive frequencies of this series are emitted when the valence-electron falls from the second, third and consecutive orbits of this succession into the first one. One would suppose that the valence-electron normally abides in this first orbit. But if this were so the energy required to ionize the atom would be $9.609h \cdot 10^{14}$, equivalent to 3.96 volts; and waves of the frequencies given by (D) could displace the electron and be absorbed thereby. But the ionizing-potential of the atom is about 25 volts and the frequencies (E) do not appear as dark lines in the absorption-spectrum of helium. Therefore there must be still another orbit much deeper down, with a much higher (negative) energy-value, than any listed under (F). In 1924 22 Lyman discovered (with his highly-curved grating and shortened light-path, and pumping arrangement for keeping the pressure low) a new series of lines of wave-lengths 584.4A, 537.1A, 522.3A and 515.7A. Their frequencies are

(G) 51.34, 55.85, 57.44, 58.18

which may be written as the succession of numbers

(H) $(59.49-8.15)$, $(59.49-3.64)$, $(59.49-2.05)$, $(59.49-1.31)$.

Comparing these with the succession (E) we recognize the same set of subtrahends,[14] and accordingly identify the common quantity $59.49 \cdot 10^{14}$

[13] It is customary to designate the orbits by their energy-values divided by hc, or $19.68 \cdot 10^{-17}$.

[14] It would not be necessary to call attention to this if we could calculate the frequency of the series-limit, which would give the energy-value of the new orbit immediately; but the four discovered lines are hardly sufficient for such an extrapolation (there are fourteen of the other series to use for calculating its limit).

as $-1/h$ times the energy-value of an additional orbit. If this orbit is the permanent home of the valence-electron, the energy required to ionize the atom must be $+59.49h \cdot 10^{14}$, equivalent to 24.5 volts. When the new lines were discovered, the accepted value was 25.3 volts, largely because of a certain measurement by Franck. After the publication of Lyman's discovery, Franck re-examined his method and data and found them compatible with the value 24.5 volts; and very recently C. A. Mackay has ascertained that the ionizing-potential of helium is 14.1 volts greater than that of mercury, which is quite definitely known to be 10.1. One could hardly desire a better illustration of the confluence of measured values of the energies of atoms and measured values of their radiation-frequencies, when both are interpreted according to the contemporary theory of radiation.

These newly-discovered waves must be the shortest in the spectrum of helium; the atom cannot emit a ray of wave-length less than the series-limit 504A, calculated by the equation

$$h\nu = hc/\lambda = \text{energy-value of the deep-lying orbit} = 59.49 \cdot 10^{14} h.$$

They are much shorter than the waves of the Lyman series of hydrogen, which Bohr's theory, together with the observed value of ionizing-potential of atomic hydrogen, justify us in declaring to be the shortest waves emitted by that atom. Furthermore, it is almost certain that they are shorter than any waves for which the valence-electron of any atom is responsible; for the ionizing-potential of helium is greater than any other measured ionizing-potential, and there is no reason to believe that any of the yet unmeasured ones exceed it. Its nearest rivals are the ionizing-potentials of the inert gases which share the last column of the periodic table. The experimentalists have not agreed very well in their estimates of these, although all agree that the values are comparatively high. Hertz, the latest to make measurements upon neon and argon, gives 21.5 volts for the first and 15.3 volts for the second. Both, therefore, should emit some rays lying below 1200A, but above 575A and 800A, respectively, and resulting from transitions of the valence-electron. Dejardin gives 12.7 volts for the ionizing-potential of krypton and very lately 10.9 volts for that of xenon. In the other columns of the periodic table, the values of ionizing-potential are prevailingly lower than in the column of inert gases. The value 10.1 volts (for mercury) is the highest among the metals; several of the non-metallic elements appear to have ionizing-potentials between 12 and 17 volts, but for some of these it is difficult to tell whether the observed value pertains to the atom

or to a molecule. The experimental material is abundant[15] enough to give practical certainty that "valence-electron rays" below 1000A occur in the spectra of only a few elements, and below 500A in none.

Nevertheless, Millikan and Bowen, photographing the spectra of all of the first twenty elements (neon and argon excluded, and chromium and copper added) down to the extremity of the region accessible with the concave grating, discovered great numbers of lines, of which they attribute dozens or scores to particular elements (for example, some forty lines ascribed to potassium, though its ionizing-potential of four volts corresponds to a minimum wave-length exceeding 2500A). Some of these may be lines of compound origin, resulting from two simultaneous changes in the electron-system of the atom, one being a transition of the valence-electron and the other a rearrangement of the other electrons. (Saunders mentions such lines in the spectra of elements of the second column of the table.) Others are due to rearrangements of internal electrons following upon a displacement of one of these. Many others are attributed to displacements of the valence-electrons of ionized atoms. Of this new field of research, the spectroscopy of ionized atoms, I wrote briefly in the first article of this series. In the more easily accessible regions of the spectrum, Paschen had discovered rays of doubly-ionized aluminium and Fowler rays of trebly-ionized silicon.[16] Millikan and Bowen go a step further by identifying certain rays of quadruply-ionized phosphorus; indeed they believe that, under the violent excitation provided by their vacuum-spark, the waves emitted by atoms which have lost all but one of the electrons from their outermost electron-shells (the three just specified are in this state) are especially abundant and intense.

[15] In the periodic table of Fig. 3 the ionizing-potentials of the elements are given along with their atomic numbers. Overlined figures are values calculated from series-limits and confirmed by direct experiment; starred values are data of experiment, for elements of which the series have not been worked out; the remaining values are calculated from series-limits and have not been verified. The data are from the cited sources, from Foote and Mohler, and from Saunders' graphic tabulation (*Science*, volume 56, pp. 50-51, January 18, 1924). Interesting variations with atomic number are observed which yield bases for estimating the values for still other elements by interpolations.

[16] "It might be mentioned that the spectrum of silicon was first selected for investigation on account of its astrophysical interest. The lines representing successive stages of ionization of this element appear in stars which there is every reason to believe are at successively higher temperatures. The complete series-data for the four spectra [trebly-, doubly-, once- and non-ionized atoms] and the ionization-potentials deduced from them, may be expected to find an important application in fixing the scale of stellar temperatures . . . All the series predicted in [these four] spectra of silicon; in the spectra of doubly-, once-, and non-ionized aluminium, of once-ionized and neutral magnesium, and of neutral sodium, have been actually produced and have been found to have the character and constants expected."—A. Fowler.

The "valence-electron" rays emitted by ionized atoms should lie at lesser wave-lengths (roughly ¼ as great) than the valence-electron rays of neutral atoms, and therefore should be particularly at home in the region newly opened to exploration. The highest frequencies emitted by the ionized-helium atom are perfectly calculable from Bohr's theory; they are the frequencies of the Lyman series, quadrupled, and the wave-lengths therefore lie between 304A and 230A. They have not been reported, but Lyman in 1919 observed two lines in the spectrum of violently-excited helium, near the positions 1214.9 and 1640.4A calculated for the first and third lines of the next helium series (having the frequencies of the Balmer series, quadrupled). The place of the second member of the series was obscured by an alien line.

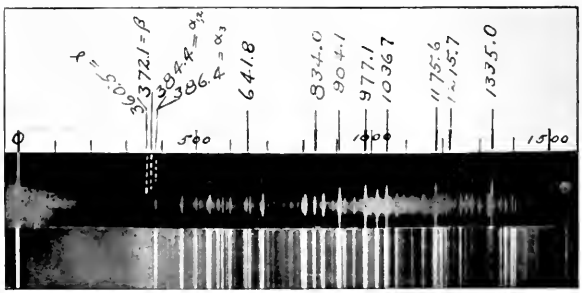


Fig. 4—Spectrum of a vacuum spark between carbon electrodes. (*Astrophysical Journal.*)

The once-ionized lithium atom, judging from the example of neutral helium, should display higher frequencies than any other once-ionized atom, and they should be arranged in recognizable series, somewhere near the extreme limit of the explorable range as it stands at this moment. They have not, however, been reported; Millikan says that his plates show no lithium lines of any sort from 1700A down at least to 370A, if not farther.

As an example of a spectrum extending far into the newly-conquered field, a plate representing the spectrum of a vacuum spark between carbon electrodes is reproduced from one of Millikan's articles as Fig. 4. The actual spectrum is in the middle; it is drawn out for better intelligibility, at the side. Most of the marked lines, including the extreme line at 360.5A and the strongest line at 1335A, are attributed to carbon; some to other elements, particularly the

one at 1215.7 which is the first line of the Lyman series of hydrogen. The interpretation of spectra like this is not a simple matter of putting electrodes of the desired substance into the tube and ascribing to it all the lines which come out on the plate. It appears that impurities, even when present in what might be considered small proportions, contribute their own rays to the spectrum in great abundance and intensity. Millikan found that all the lines present in the spectrum of the vacuum spark between magnesium electrodes were also present when aluminium electrodes were used, and vice versa, and finally assigned them all to oxygen. Lyman found it extremely difficult to decide which lines belong to hydrogen and which to helium, since the spectra of glow-discharges in these gases have so many lines in common. Helium has a pronounced habit of encouraging excitation of the rays of whatever other gases are mixed with it, since the helium atoms require so much energy to displace their valence-electrons that free electrons shot into helium gas are liable to bounce harmlessly from one helium atom to another until they strike and excite an atom of another variety. Even if one can be sure that all the rays in a spectrum belong to a single element there remains the problem of assigning them to neutral or variously-ionized atoms. It is clear that the completion of the spectroscopist's task is deferred by this extension of it to what the Germans call the unforeseeable time.

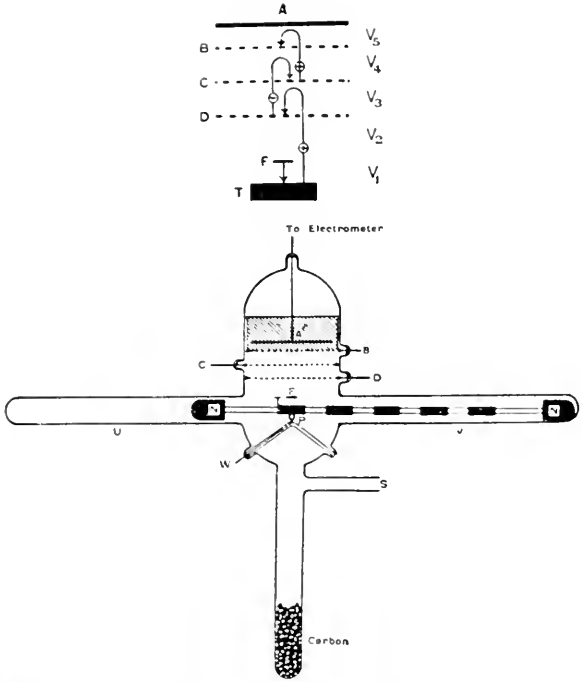
We return to the consideration of the lacuna in the spectrum, which extends from 13A up to a boundary which by the use of high vacua, concave gratings, and violent excitations, has been forced from 1200A down to 136A. This wave-length 136A stands for the moment as the lowest which has ever been actually measured with the ruled grating; and in spite of the unexpected and fortunate adequacy of the instrument down even to this point, little more can be demanded from it. The reason is, that the substance on which the rulings are made must eventually cease to reflect the rays on account of its own looseness of texture. Being a congeries of atoms themselves separated by finite distances, the metal will not behave as a continuum towards waves of a length not very large compared to its own atomic spacing. Below 13A waves are not reflected. Little is known of the rate at which the reflecting-power dwindles away to zero between 136A and 13A and this little we owe again to Holweck. He directed a beam of radiation (it was a mixed beam, as was previously made clear, and the wave-length-value is merely the minimum wave-length in it) against a polished bronze mirror at the very oblique incidence of 73.°8; the reflected beam had one-third the intensity of the incident beam at wave-length 123A (practically the extreme wave-length of Millikan's

experiments), but only 10% at 60A and only 3% at 40A. The performance was much better at a still more oblique incidence; on the other hand Holweck thinks that it gets rapidly worse as the incidence is made more nearly normal, and if this is true the outlook for the concave grating, with its condition of almost normally-incident light, is most unpromising.

Below the boundary 13A, the atomic constitution of solid substances turns from a hindrance into an advantage, and crystals serve as natural diffraction-gratings of incomparable fineness—too fine, indeed, for our convenience in this part of the spectrum, since the boundary is fixed by the smallness of the distance d between successive layers of atoms in the diffraction-grating. Rocksalt, one of the standard crystals, for which $d = 2.814$A, has been used successfully up at least to 1A (by Fricke) and the rest of the way to 13A has been explored with crystals of gypsum ($d = 7.58$A) or sugar ($d = 10.56$A); in this region it was necessary to evacuate the light-path, precisely as in the region beyond 136A. The only possibility of a new advance depends on the utilization of crystals of still greater inter-atomic spacings. Holweck mentions a crystal with a formidable name, for which $d = 19$A, and de Broglie and Friedel found that the oleates of sodium, potassium and ammonium presented spacings of the order of 40A between consecutive molecule-layers. If these substances can be adapted for use in crystal spectographs, the boundary of the explored region may be pushed far beyond its present place. It may be found, however, that the crystal absorbs the rays before they go deeply enough to be diffracted.

As for the region between 13A and 136A, no one has ever measured the wave-length of a radiation lying within it; but there is a method which indicates the existence and something about the wave-lengths of rays which almost certainly belong in it. In applying this method the photographic plate is replaced by a metal electrode (usually of platinum) which, when irradiated by rays of any wave-length (less than a certain critical one which always lies far above this range) emits electrons. For this reason it is often known as the *photoelectric method*, although the substitution of one kind of receiver for another is not its most distinctive characteristic. A target made of the substance to be studied is sealed into a tube, opposite a source of electrons (generally a filament for thermionic emission); the photosensitive electrode is placed somewhere in the tube where whatever rays are excited at the surface of the target will fall directly upon it. It is all-important to protect this electrode from electrons and ions, negative or positive, proceeding from target, filament, or anywhere else.

Usually the electrode is screened by a family of gauzes, with their potentials adjusted as is indicated in Fig. 5 (with the paths of intruding ions of both signs, including those excited from the outer gauzes themselves, mapped out to show how they are rebuffed). Naturally



Figs. 5 and 6. Horton's apparatus for determining excitation-potentials by the "photoelectric method." (*Philosophical Magazine.*)

the arrangement of potentials in front of the electrode must be such that the emitted electrons are all drawn away from it, not driven back onto it. The rate of emission of electrons, the *photoelectric current*, may be measured with an electrometer connected either to the sensitive electrode or to a gauze so placed as to gather in all the electrons emitted from it. Figs. 5 and 6, the latter of which shows a

completely equipped tube with filament at F, a row of targets which can be moved consecutively into place at T, and the photosensitive electrode at A with its gauze shields in front of it, come from the work of Horton, Andrewes and Davies. A tube designed and used by E. H. Kurth is shown in Fig. 7; the filament is seen in perspective at C, the target at T, and the sensitive disc at D; the family of diverging straight lines represents a set of metal laminae, which being charged

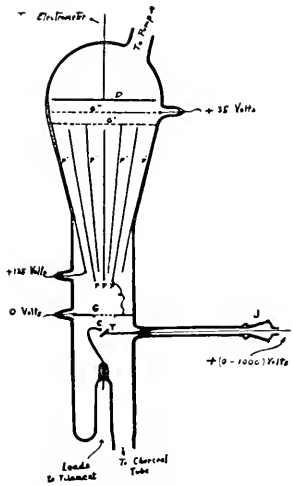


Fig. 7 Kurth's apparatus for determiniog excitation-potentials. (*Physical Review.*)

alternately to potentials 0 and +135 volts gather in any ions which start up towards the disc. The method can also be adapted to gases, and this application has an interesting and important history; but as nearly all the data respecting gases refer to wave-lengths superior to 1200A, they fall out of the province of this discourse. Foote and Mohler, however, penetrated to 26A with the apparatus of Fig. 8, filled with oxygen. The filament is at A; the electron-accelerating voltage V is applied between A and the gauze B, so that the target is essentially a thin layer of gas enveloping B; the photosensitive electrode is the gauze C, the photoelectric current from which is gathered in by the plate D (screened against positive ions by its high potential).

The art of detecting radiations by this method consists in giving various values to the "bombarding voltage" V between target and filament, which is the measure of the energy of the electrons impinging on the target; measuring the photoelectric current i, which is the measure of the intensity of the rays; plotting i (or better the ratio of i to the current of bombarding electrons) versus V; and examining the curve to see whether it displays sudden changes of slope. If it does,

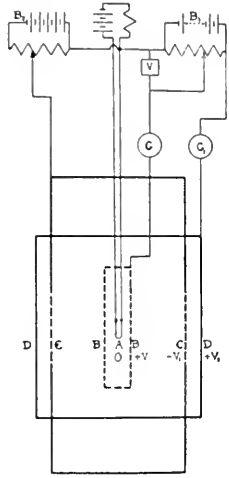


Fig. 8.—Mohler and Foote's apparatus for determining excitation-potentials of gases and vapours. (*Bulletin of the Bureau of Standards.*)

one infers that at the corresponding voltages new radiations suddenly burst forth. The method therefore consists in finding critical bombarding-voltages, that is, critical electron energies which just suffice to excite particular sorts of radiation; it is a method for discovering *excitation-potentials*. Three excellent instances of such abrupt changes in slope, or *breaks* as they are frequently called, appear in the (i, V) curve determined with an aluminium target by Horton and his associates (Fig. 9). Very many such curves appear in the literature, with more or less conspicuous breaks; some are as striking as these in the figures, some require a good deal of care and experience to locate them properly, and some, one is driven to conclude, are visible only to the eye of faith. But it is hardly possible to doubt that such a

corner as the three here reproduced marks the *entrée en scène* of a new ray or set of rays.[17]

But a determination of an excitation-potential is not a measurement

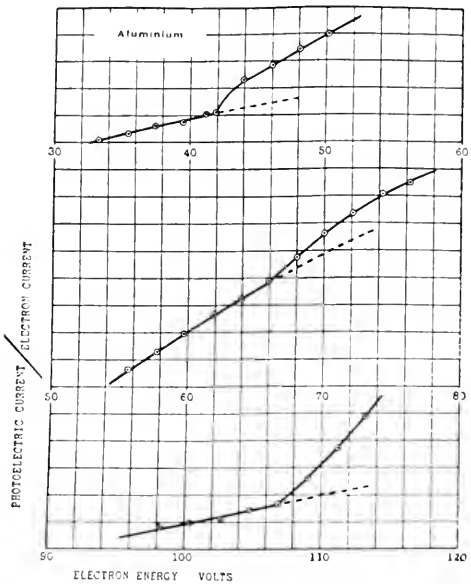


Fig. 9—Breaks in a photoelectric-current curve indicating excitation-potentials. (*Philosophical Magazine.*)

of the wave-lengths of the excited rays; and while it is supposed that excitation-potentials between 1000 volts and 100 volts are associated with rays of wave-lengths between 12A and 123A, this is merely a supposition.[18] We require a theoretical relation between excitation-

[17] It is clear from this account that the photosensitive disc might be replaced by a photographic plate, on which the opacity due to the rays produced by consecutive values of V could be measured, or by an ionization-chamber, in which the ionization-currents could be measured. It is equally clear that neither method would be so suitable for detecting slight discontinuities in rate of increase of radiant energy with increase of V. However, both methods are used at higher frequencies, where by dispersion of the waves a discontinuity in the intensity at a single wave-length is made more conspicuous.

[18] This is the best place to remark that electrons of voltage V bombarding a solid, in addition to exciting (if V is high enough) rays characteristic of the bombarded atoms, excite also a continuous spectrum of rays of all frequencies up to a maximum

potentials and excited frequencies. The question is of high importance, not simply because we are interested to know whether some of the excited rays really lie in the hitherto unpenetrated range, but primarily because excitation and emission are among the fundamental qualities of atoms. Excitation-potentials exceeding 1000 volts generally produce rays of which the wave-lengths are less than 12A and can be measured with the crystal spectrograph, so that a rule or law can be deduced from the two sets of measurements. Excitation-potentials inferior to 25 volts generally produce rays of which the wave-lengths are greater than 500A and can be measured with optical apparatus, and again a law can be deduced from the two sets of data. But the law is not the same in the two cases; this is because excitation, in the former case, consists in displacing a deep-lying electron, while in the latter case it consists in a displacement of the valence-electron. We are forced to the disconcerting conclusions that excitation-potentials between 1000 volts and 25 volts involve electrons of an intermediate type, and that the still-unverifiable law connecting them with the frequencies of their excited rays is not identical with either of the laws in the accessible regions of the spectrum.

The law for excitation-potentials involving displacements of the valence-electron is twofold. Each atom has at least two such excitation-potentials. One of them is its ionizing-potential. When the accelerating-voltage of an electron-stream playing against a multitude of free atoms forming a gas is raised just past the value V_i at which an individual electron has just enough energy to remove the valence-electron of an atom, there is an outburst of radiation. This comprises rays of many frequencies probably all those which we have called valence-electron rays and they are emitted as the valence-electrons descend step-by-step along their ladders of orbits. All these frequencies conform to the relation

$$(1) \qquad h\nu < eV_i.$$

The other excitation-potential is the *resonance* potential of the atom (there may be more than one of these [19]). When the accelerating-

equal to eV/h. The heterogeneous beams used by Holweck in the experiments previously cited consisted chiefly, if not entirely, of this continuous spectrum. All the excitation-potentials mentioned in these pages, however, relate to individual rays or groups of individual rays characteristic of atoms.

[19] This question is still incompletely solved, in spite of much labor. At one time it was supposed that the valence-electron could be raised either altogether out of the atom, or else to the deepest-lying of the transient-sojourn orbits or to either of the two deepest-lying orbits, if there are two complete families of orbits such as the mercury atom possesses; but not to any of the other transient-sojourn or "virtual" orbits. This restriction would apply only to displacements caused by impinging electrons; quanta of appropriate frequencies can lift the valence-electron to any of

voltage of the electron-stream is raised just past the value V_r at which the individual electron has just enough energy to raise the valence-electron from its normal to one of its transient-sojourn orbits, there is an outburst of radiation This comprises rays of a single frequency, emitted when the valence-electrons return in single leaps from the orbits to which they were momentarily raised to the orbits of their normal habitation. This frequency conforms to the relation:

$$h\nu = eV_r. \tag{2}$$

The law for excitation-potentials involving displacements of deep-lying electrons bears a certain resemblance to the first of the foregoing laws. When the accelerating-voltage of an electron-stream playing against a multitude of atoms assembled in a solid or liquid is raised just past the value V_e at which an individual electron has just enough energy to extract a certain deep-lying electron, say a K-electron, there is an outburst of radiation comprising many frequencies, all conforming to a relation resembling (1), to wit:

$$h\nu < eV_e. \tag{3}$$

But it would be misleading to assume that the processes resulting in (1) and in (3) are identical. In the first place, it is not certain that the deep-lying electron need be completely extracted. Suppose it possessed a set of transient-sojourn orbits in the outskirts of the atom, their energy-values differing from one another and from that of the "orbit at infinity" (the state in which the electron is quite detached) by amounts less than the 25 volts which is the maximum difference between the energy-values of any valence-electron. Then there might be several excitation-potentials, differing from one another by 25 volts at most; but this difference would be so inconsiderable a fraction of the value of the extraction-potential V_e which ranges from more than 100,000 volts for the K-electrons of uranium to about 1100 volts for those of neon, that they would be difficult to distinguish. Indications of multiple excitation-potentials have, how-

an immense number of orbits of a certain set, but not to transient-sojourn orbits of certain other sets. Lately it has been affirmed that impinging electrons of the right energy can lift the valence-electron to any one at all of its transient-sojourn orbits, even those to which it cannot be lifted by quanta; but this rule, if it is the true one, has not yet been illustrated by any extensive set of experimental data, though Hertz has lately intimated in a brief note that he has assembled such a set by experiments on helium. Franck and Knipping detected excitation-potentials corresponding to the lifting of the valence-electron of helium from its normal orbit to several distinct P-orbits; but I gather from a later paper by Franck that nobody has been able to reproduce the result. Olmstead and Compton discerned excitation-potentials corresponding to the lifting of the electron of hydrogen from its normal orbit to each of the next six transient-sojourn orbits.

ever, been discerned in the "fine structure" of the K absorption-edges of the lighter elements (notably the elements from sodium to potassium). In the second place, the process of emission is different in the two cases described by equations (1) and (3). In the former case, the rays were emitted as the valence-electron (or another replacing it, which comes to the same thing) redescended its ladder of orbits; but when a deep-lying electron is extracted, the resulting rays are emitted because of rearrangements of the other internal electrons of the atomic electron-system, which occur irrespective of whether the departed electron quickly returns to the atom, or remains a long time away.

I will now risk the making of a distinction which may eventually turn out not to be the most natural or practical, by reserving the name *deep-lying electrons* for those electrons which lie entirely within at least one completed electron-shell of an atom, and designating the others (exclusive of the valence-electron, which has already been set apart from the rest) as the *shallow-lying electrons*. It follows from this definition that the first nine atoms of the periodic table, up to fluorine (inclusive) possess only shallow-lying electrons; the next eight (Ne to Cl) have one set of deep-lying electrons, the K set; the next eighteen (A to Br) have at least four sets of deep-lying electrons, the K set and three L-sets (the last three can be grouped as one). It follows also that every instance in which an excitation-potential has been measured, and the wave-lengths of the excited rays have also separately been measured, is an instance in which a deep-lying electron is involved. For example, the excitation-potentials involving extraction of the K-electrons have been measured from the top of the periodic table down to the twelfth element (Mg), over which range they decline from 115,000 volts to 1100 volts; the excited waves have been measured over the same range and down to the eleventh element (Na), over which range they rise (for the principal ray) from .10A to 11.88A. At this point, and just before the K-electrons pass over into the category of shallow-lying electrons at the ninth element, the wave-lengths enter into the inaccessible range. The wave-lengths of the rays excited when one of the L electrons is displaced have been measured from the top of the table down to the twenty-ninth element (Cu) where, arriving at 13.3A, they too pass into the immeasurable class.

The general consequence of all this is, that the excitation-potentials involving shallow-lying electrons must be below 1000 volts; that, conversely, the excitation-potentials observed between 25 volts and 1000 volts are chiefly those of excitations which consist in displace-

ments of shallow-lying electrons; and finally, that the wave-lengths of the excited rays lie below 13A, many of them in the inaccessible range, some in the range newly opened to exploration. This is a most unfortunate coincidence, for instead of being able to apply laws which prevail in other ranges to compensate for our inability to measure wave-lengths in this range, we have to expect distinct laws within it. Must shallow lying electrons be extracted altogether from the atom if they are to be displaced at all or have they certain transient sojourn orbits to some or all of which they may be raised by electron-impacts? Do the emitted rays result from a step-by-step return of the displaced electron? or from a return in a single leap? or from a rearrangement of the remaining electrons? or from a compounding of changes of the two latter types? So long as the emitted wave-lengths are not measured, these questions cannot be answered with confidence.

Some little can be inferred from numerical relations among excitation-potentials. McLennan and Clark, for example, observed three excitation-potentials of lithium, at 37.0, 31.8 and 12.0 volts. The first two of these voltages stand nearly in the ratio of the first two frequencies of the Lyman series in the hydrogen spectrum, which suggested to the discoverers that the processes involved in the excitations were the raising of a K-electron to the first and second of a pair of transient-sojourn orbits, standing in the same relation to the normal orbit of the K-electron as the orbits of energy-values $-Rh\ 4$ and $-Rh\ 9$ stand to the normal orbit of energy-value $-Rh$ in the hydrogen atom. That is to say, they conceive these excitation-potentials to be comparable to resonance-potentials, and the K-electron of lithium to behave like a valence-electron. They also found excitation-potentials of beryllium at 20.3 and 16.0, and of boron at 27.92 and 23.45. The ratio of each pair of numbers is about equal to the ratio of the first two frequencies of the Balmer-series, suggesting that these are resonance-potentials of an L-electron; the details of the analogy may be left to the reader to work out. Each of the latter elements displayed additional higher potentials, to be associated with the K-electrons. Rollefson lately discovered seven excitation-potentials of iron in the range between 160 and 264 volts, expressible by a formula $(a-b\ n^2)$ if the integer values 5, 6, 7, 8, 9, 10 and 12 are successively given to n. If these seven potentials correspond to elevations of a certain shallow-lying electron to seven transient-sojourn orbits, the extraction-potential for this electron can be calculated by an extrapolation (so also in the cases cited from McLennan and Clark). Rollefson interprets certain other excitation-

potentials as corresponding to elevations of certain deep lying electrons to transient-sojourn orbits.

Some assistance in identifying the excitation-potentials of the light atoms can be obtained by plotting the recognized excitation-potentials of the heavier atoms, and also the frequencies of the rays excited; plotting curves representing them as functions of atomic number; and extrapolating the curves into the range of low atomic numbers. The best procedure is to plot the square roots of the excitation-potentials and the emission-frequencies, as then the curves are nearly straight lines (Moseley's law). Some of these lines are shown in

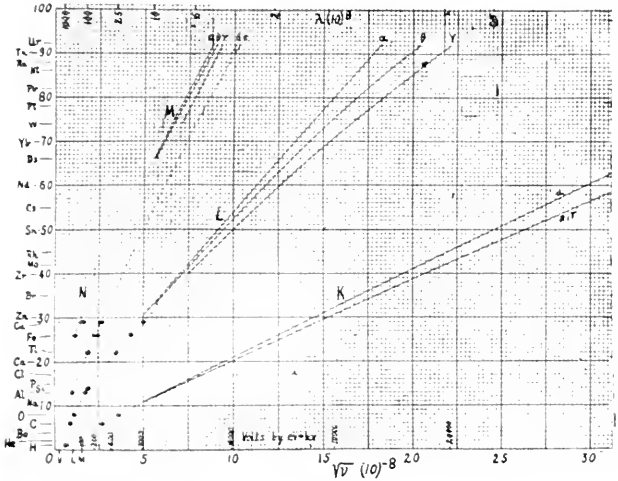


Fig. 10. Curves representing square roots of emission-frequencies of heavier atoms as functions of atomic number. (*Physical Review.*)

Fig. 10 (from Kurth). Since the atomic numbers are laid off (contrary to usage) along the axis of ordinates, the lowest-lying line represents the highest recognized emission-frequencies (the $K\delta$ and $K\gamma$ frequencies, which actually are slightly different, but are not indicated separately upon the graph). The next line, marked $K\alpha$, represents another particular emission-frequency. Excitation is the same for every ray of this group, and consists in extracting one of the deepest-lying or K electrons of the atom; and the excitation-potential for the

entire group, the K excitation-potential, is also represented by a straight line, the K line, which may be taken as coincident with the lowest-lying line in the graph, provided that we translate frequencies into potentials by the relation $V = h\nu/e$ (both frequencies and

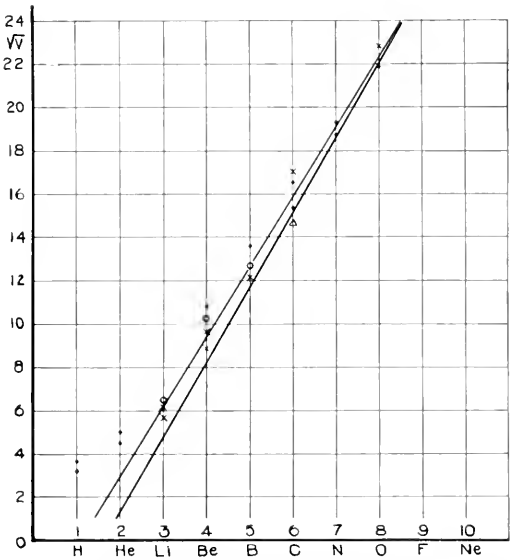


Fig. 11—Excitation-potentials of light elements, correlated with displacements of K electrons. (Cf. footnote 20.)

potentials are laid off along the axis of abscissæ). This K line, it must be realized, extends the whole way from atomic number 92 to atomic number 12.

The circles upon the graph represent excitation-potentials inferior to 1000 volts, observed by Kurth. Three of these lie very close to the downward prolongation of the K line; the almost inevitable inference is, that in these three cases the excitation consists in the extraction of one of the electrons nearest the nucleus. The others lie so much above the extended K-line that they must belong to a distinct class. Many additional measurements have been made

since this graph was published, and in Fig. 11 I have set down all the experimental values known to me which have been given for excitation-potentials of the first eight elements, omitting those which are so small that they obviously do not belong to the K class.[20] The

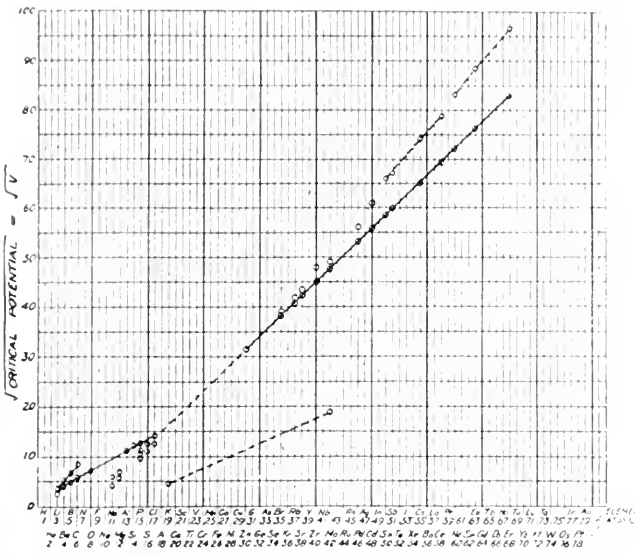


Fig. 12—Emission-frequencies of heavier and excitation-potentials of lighter elements, correlated with L electrons. (*Proceedings of the Royal Society.*)

[20] The data are from various sources, as follows. The dots for hydrogen and helium represent the observed ionization and resonance potentials of these atoms. The dots for *Be*, *B*, *C*, *N* and *O* are at values of excitation-potentials given by Mohler and Foote from experiments on gaseous compounds of these atoms. All the other data except Holweck's are values of excitation-potentials for solids. The crosses for *Li* and *Be* stand for the excitation-potentials observed by McLennan, the circles for the extraction-potentials of the K electrons which they infer from these data. The cross for *B* represents three values lying so close together as to be indistinguishable (from McLennan, Hughes, and Holtsmark) and the cross for *C* also three coincident values (Kurth, Richardson and Bazzoni, Holweck). The circle for *B* is at the potential corresponding to a discontinuity in absorption, observed by Holweck. The triangle for *C* is a value observed by Hughes, and the cross for *O* a value from Kurth (obtained with oxidized copper). No data for *F* or *Ne* are available. At *Na* measurements on the wave-length of $K\alpha$ and at *Mg* measurements on the K absorption-edge commence.

lower of the continuous straight lines coming downward from the right is the prolongation of the K line from the heavier elements downward; the upper is the prolongation of the $K\alpha$ line. The fact that these intersect proves that linear extrapolation from the range of heavier atoms is unjustifiable.

Reverting to the graph in Fig. 10, the problem of properly extra-

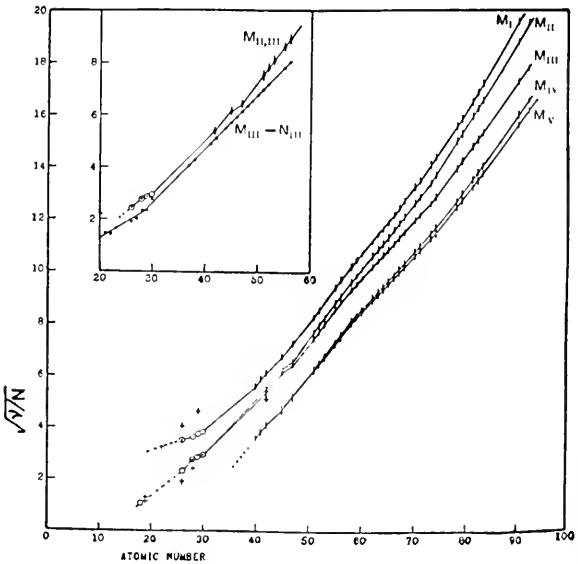


Fig. 13 Excitation-potentials correlated with M electrons. (*Philosophical Magazine.*)

polating the L and M curves is clearly not so simple as it was for the K curve; since they extend over shorter segments and do not come so far down into the range of light elements. In Fig. 12 (from McLennan and Clark) the circles for the elements from number 3 to number 17 represent observed excitation-potentials which they attribute to displacements of L electrons; those for the elements from number 30 to number 69 represent the highest and the lowest recorded emission-

frequencies of the L-series for these elements, frequency being translated into equivalent voltage by the same relation as above. As for the excitation-potentials of the heavier elements, few measurements on potentials of the L class have been made, and very few indeed upon potentials of the M class—not nearly enough for an extrapolation. The deficiency is partially compensated by calculating the L and M excitation-potentials from the K, L and M emission-frequencies—an elaborate process, requiring a good deal of care in measuring and properly interpreting the various emitted rays. In this manner the potentials for the group of five M levels have been estimated for the various elements from the ninety-second down to the fortieth, and Horton has attempted to link onto them certain excitation-potentials which he and others observed when bombarding elements between number 20 and number 30 with electrons (Fig. 13). The curves must be supposed to bend, somewhere between the thirtieth and the fortieth elements; it is in this region that the M electrons pass from the status of deep-lying to the status of shallow-lying electrons. The excitations and emissions involving the shallow-lying electrons of the heavy atoms form a complicated system, of which the study has scarcely been begun, and will certainly prove perplexing. When research in this field is completed, each of the excitation-potentials and each of the emission-frequencies of every kind of atom will be entered upon curves, each of the curves corresponding to a definite and definitely-pictured process of rearrangement in the atomic electron-system, and extending over all the atoms of the periodic table which can be theatres of that process. This achievement may be reserved for a later generation.

LITERATURE

F. S. Brackett: Phys. Rev. 20, pp. 111–112; 1922.

British Association symposium "Spectra of the Lighter Elements"; Nature 112, pp. 217–224; 1924.

M. de Broglie and G. Friedel: C.R. 176, pp. 738–740; 1923.

G. Dejardin: C. R. 176, pp. 894–897; 1923. ibid. 178, pp. 1069–1071; 1924.

P. D. Foote and F. L. Mohler: *Origin of Spectra* (Chemical Catalog Co., 1922).

J. Franck: ZS. f. Phys. 11, pp. 155–160; 1922.

H. Fricke: Phys. Rev. 16, pp. 202–215; 1920.

G. Hertz: ZS. f. Phys. 18, pp. 307–316; 1923. Naturwiss. 11, pp. 778–779, 1923.

J. Holtsmark: Phys. ZS. 23, pp. 252–255; 1922.

F. Holweck: Annales de Physique, (9) 17, pp. 5–53 (1922). C. R. 173, pp. 709–712; 1922. (C. R. 176, pp. 570–573; 1923 (reflection of X-rays).

J. J. Hopfield: Phys. Rev. 18, p. 327 (1921), 20, pp. 573–588 (1922).

F. Horton, C. Andrewes, A. C. Davies: Phil. Mag. 46, pp. 721–741; 1923.

A. L. Hughes: Phil. Mag. 43, pp. 145–161, 1922.

E. H. Kurth: Phys. Rev. 18, pp. 461–476; 1921.

T. Lyman: *The Spectroscopy of the Extreme Ultra-violet* (Longmans, Green & Co., 1914). Astroph. Jl. 43, pp. 87–102; 1916. Nature 104, p. 314; 1922 (lines of ionized helium). Phil. Mag. 41, pp. 814–817 (1921) and Science, 56, pp. 167–168 (1922) (lines of neutral helium; with H. Fricke,) J. Opt. Soc. 7, pp. 495–499; 1923. (Vacuum grating spectrograph.)

C. A. Mackay: Phys. Rev. 21, p. 717; 1923.

J. C. McLennan: Proc. Roy. Soc. 95A, pp. 258–273; 1919 (with R. J. Lang). ibid. 98A, pp. 114–123; 1920. (Vacuum grating spectrograph). ibid. 102A, pp. 389–410; 1923 (with M. L. Clark; excitation potentials).

R. A. Millikan: Astroph. Jl. 52, pp. 47–64; 1920. ibid. *53*, pp. 150–160; 1921 (with R. A. Sawyer, I. Bowen). Phys. Rev. 23, 1–34; 1924 (with I. Bowen). Cf. also R. A. Sawyer.

F. L. Mohler and P. D. Foote: Bull. Bur. Standards 17, pp 471–496; 1923.

P. S. Olmstead and K. T. Compton: Phys. Rev. 22, pp. 559–565; 1923.

O. W. Richardson and C. B. Bazzoni: Phil. Mag. 42, pp. 1015–1019; 1921.

G. K. Rollefson: Phys. Rev. 23, pp. 35–45; 1924.

F. A. Saunders: Astroph. Jl. 40, pp. 377–384; 1914. (Vacuum arc spectra.) Science 59, pp. 50–51; 1924. (Ionizing potentials.)

R. A. Sawyer: Astroph. Jl. 52, pp. 280–300; 1920.

F. Simeon: Phil. Mag. 46, pp. 816–819; 1923. Proc. Roy. Soc. 102A, pp. 484–496; 1923. ibid. *104A*, pp. 368–375; 1923.

A. von Welsbach: Ann. d. Physik 71, pp. 7–11; 1923.

An Electrical Frequency Analyzer[1]

By R. L. WEGEL and C. R. MOORE

Synopsis—An apparatus has been developed by means of which it is possible to measure and obtain a permanent record of the frequency components of an electric current wave. The device has two frequency ranges: 20 to 1250 cycles and 80 to 5000 cycles; the amount of power required does not in general exceed 500 microwatts; and the time necessary for making a record is about 5 minutes. An attachment is provided which permits of the making of simultaneous harmonic analyses of two complex waves in the same length of time.

In principle, the process consists in feeding the complex wave to be analyzed into a selective network, the essential feature of which is a sharply tuned circuit whose frequency of tuning is controlled by varying the capacitance in small steps with a pneumatic apparatus similar to that in a player piano. A maximum of response of the circuit occurs at each frequency of tuning which coincides with a component of the complex wave. An automatic photographic recorder of the response to each frequency of tuning is provided by means of which the frequency and magnitude of each component of the complex wave may be obtained. For convenience of operation, an automatic control apparatus is provided, so that it is only necessary to connect the complex source or sources to be analyzed and press a starting button. The completed record of the analysis is delivered after the machine has passed through the entire range of frequencies.

The application has so far been principally to problems in the communication field such as the analysis of performance and distortion at audio frequencies of vacuum tube and mechanical oscillators and amplifiers, analysis of complex telephone waves and speech sounds, and the effect on a complex wave of transmission through electrical and acoustic apparatus. In the power field many applications are obvious, such as for example, quantitative comparison as to frequency content of the voltage and current supplied to and delivered by transformers, voltage and magnetic flux studies in generators and motors, commutation, and the effect of wave-shapes in power transmission line problems and control apparatus.

Introduction

THE harmonic analyzer described in this paper consists of a variable tuned circuit into which the complex current wave to be analyzed is introduced, and an automatic recording apparatus to register its response as the frequency of tuning is changed.

The first recorded use of a tuned circuit as an analyzer was by Pupin in 1894.[2] He analyzed power waves by measuring the response of circuits tuned to each of the harmonic frequencies. It has been the practise for a number of years to determine the frequency characteristics of currents and voltages on power circuits and noise on telephone lines by means of a variable resonant circuit which includes a telephone receiver for listening.

[1] Presented at the Midwinter Convention of the A. I. E. E., Philadelphia, Pa., February 4-8, 1924.

[2] Resonance Analysis of Alternating and Polyphase Currents, Trans. A. I. E. E., Vol. XI, p. 523.

During the recent war a rapid automatic method was developed for varying the tuning of a circuit in such an analyzer in connection with the analysis of sounds radiated by submarines. The analyzer described in this paper is in principle the same as this apparatus but includes such improvements as were found desirable by experience to increase the speed, dependability and convenience of use. The present apparatus is capable of recording the frequency and magnitude of each component in a complex wave between 20 and 1250

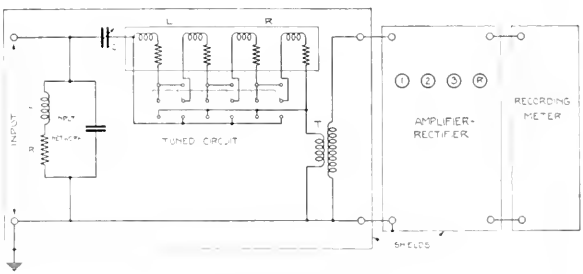


Fig. 1—Schematic Analyzer Circuit

cycles or 80 and 5000 cycles in about five minutes. This analyzer does not measure the phase of the various components but has the advantage that the frequencies need not be simple multiples of the fundamental as is the case with graphical analyzers. With this apparatus it is possible to measure quite accurately component frequencies as close together as about fifteen cycles at the lower end of the range and about 200 cycles at the upper end of the range, and to detect components as close together as three to five cycles at the lower end and fifty cycles at the upper end of the range.

Principles of Operation of the Analyzer

Fig. 1 is a schematic diagram of the essential elements of the analyzer circuit. The wave to be analyzed is introduced at the input terminals from which it passes to an input equalizing network and to the variable tuned circuit. The tuned circuit consists of a variable condenser of capacitance C and a coil whose inductance is L and resistance R. The value of the capacitance C is varied in small steps by an automatic device to be described in the next section. The inductance L consists of four identical windings on a toroidal

core which, by means of a switch, may be thrown in series or in parallel, thereby changing the value of the inductance in the ratio of 16 to 1. With the same range of capacitance values this change in inductance gives the two frequency ranges of tuning, 20-1250 cycles and 80-5000 cycles. By means of the high-ratio transformer T the response of this circuit is applied to a vacuum tube amplifier-rectifier and registered by means of the recording meter.

This circuit arrangement will analyze a complex wave by virtue of the selective shunting of current by the tuned circuit from the input network. The impedance of the source of the complex wave is in practise maintained high in value at all frequencies compared to that of the input network so that the input wave-shape is independent of the small changes in impedance of the analyzer due to the varying of condenser C. The current fed into the analyzer traverses two paths, the input network and the tuned circuit. The impedances of these paths are respectively,

$$Z_1 = \frac{(R_1 + j\omega L_1)\ j\omega C_1}{R_1 + j\omega L_1 + 1\ j\omega C_1}$$

and

$$Z = R + j\omega L + 1\ j\omega C.$$

The transformer T introduces into the tuned circuit a small resistance and inductance, both of which are negligible. The input network impedance Z_1 varies gradually from 0.4 ohms for direct current to about 10 ohms at 5000 cycles. The values of the elements are: $R_1 = 0.4$ ohms, $L_1 = 0.075$ milhenries, $C_1 =$ about 15 microfarads. Impedance Z of the tuned circuit depends on the setting of the variable condenser C. The resistance R of the iron-core coil, varies with frequency; its values for the parallel connection are 0.7 ohms for direct current, 1.5 ohms at 2500 cycles and 1.2 ohms at 5000 cycles The value of the inductance L for the parallel connection is 23.4 milhenries and is practically constant with change of frequency. For the series connection both R and L are sixteen times as great. The capacitance is varied from about 200 microfarads to about 0.05 microfarads. It will be seen that for each capacitance value there is a frequency, $f_r = 1/(2\pi\sqrt{LC})$, for which the tuned circuit impedance, Z, is R. For other frequencies Z is much greater due to the reactance. An incoming current of frequency f_r is, therefore, largely shunted through the tuned circuit while current of any other frequency passes through the input network. In this way if the capacitance C is varied gradually the tuned circuit will shunt selectively from the input network the successive components of the complex wave.

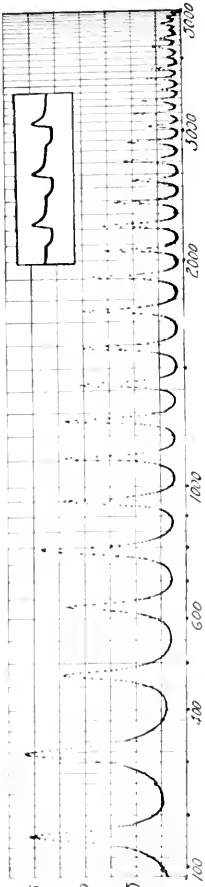


Fig. 2—Record of 160-cycle Buzzer Output

The special features of design of this analyzer circuit can be better explained by reference to a typical record made by the apparatus. Fig. 2 is the record of analysis of the current from a buzzer which vibrates with a frequency slightly under 160 per second and gives an irregularly shaped wave which is shown in the accompanying oscillogram. In taking this record the windings of the tuning inductance were in parallel so as to give the frequency range 80-5000 cycles. The vertical scale gives approximately the r. m. s. current in milliamperes at each frequency (as read on the horizontal scale) at which a peak occurs. It will be seen that a peak occurs at each multiple of the frequency of the buzzer. The r. m. s. values of input current at the corresponding frequencies as read from the peaks on the record are: 160, 1.6 milliamperes; 320, 1.6 milliamperes; 480, 1.25 milliamperes; 640, 1.2 milliamperes; 800, 1.45 milliamperes; 960, 1.25 milliamperes; 1120, 1.2 milliamperes; 1280, 1.1 milliamperes; 1440, 1.05 milliamperes; 1600, 1.0 milliamperes; 1760, 1.0 milliamperes; etc. The root square sum of all components shows that 4.7 milliamperes was the effective value of the complex current fed into the analyzer.

The fact that the 80-5000 cycle records read directly the current at each frequency component is due to the special design of the input network. A small correction is still necessary but can be neglected except where maximum obtainable accuracy is desired. If the input network were a pure resistance the higher frequency components would produce relatively lower peaks because of the falling off of efficiency with frequency of the amplifier-rectifier circuit and the increase in resistance of the tuning coil. The input network was designed empirically so that with constant input current the voltage drop across the input terminals increases with frequency in such a way as to compensate for these high-frequency losses. The tests to determine this were made by taking records of single frequencies of known amounts.

It will be seen that the frequency scale is gradually contracted as the upper end of the record is approached. Owing to the increase in resistance of the coil with frequency, the sharpness of tuning of the analyzing circuit decreases with frequency. Each peak on the record corresponding to a single frequency is a plot of the resonance curve of the variable tuned circuit. The sizes of the capacitance steps are so adjusted that a sufficient number of points, necessary to trace a resonance peak at all frequencies, is recorded. The length of the record and the time required for an analysis are determined by the number of points needed.

When peaks on the record are so close together as to overlap greatly, the reading on the scale is untrustworthy. If, instead of a rectifier and direct-current meter, an alternating-current meter giving deflections proportional to total r. m. s. values, were used, it would be theoretically possible to determine the component frequencies and amplitudes making up any composite peak, provided the number of frequencies could be determined. This procedure, however, would be impracticable. An examination of the theory of the rectifier shows that the problem of separation of the components of a composite peak is in general indeterminate. The rectifier however resolves adjacent peaks somewhat better than an alternating-current meter.

The analyzer has been most used in the analysis of audio-frequency currents for which the higher frequency range, 80-5000 cycles, is more useful. For the investigation of power problems the lower range would ordinarily be more suitable. In order to simplify the change from one frequency range to the other the tuning inductance only, is changed, leaving the mechanism for varying the capacitance in steps the same for both ranges. Since the inductance change in going from the high to the low-frequency range is in the ratio 1:16 and the change in the frequency range 4:1, the abscissas on the low-frequency records have one-fourth the value of those on the high-frequency records.

Since the smallest frequency divisions at the lower end of the high-frequency records are 20 cycles, these divisions on the low-frequency records are 5 cycles. There are, therefore, four times as many steps of tuning in the same frequency interval on the low as on the high-frequency record. The low-frequency record is therefore not of minimum practicable length. Since the same input network is used with the 20-1250 range as with the 80-5000 range, the low-frequency records are not direct reading in input current, but must be used with a calibration. Our use of the low-frequency range, however, has been so limited as not to justify the preparation of additional equipment for this use of the analyzer.

The apparatus is equipped with a device which permits of making simultaneous analyses of two complex waves. The principal reason for making such double records is to reduce errors in comparing two sources which may vary with time. The device may also be used simply to save time. It operates by connecting alternately to the analyzer the two complex waves in such a way that the record for each wave is traced by points representing alternate tuning condenser settings.

Description

The mechanism of the analyzer is so designed that to take a record it is only necessary, after starting the amplifier and connecting to a 110-volt power source, to attach the leads from the source or sources to be analyzed and press a starting button. The completed record is then automatically delivered in about 5 minutes after which the apparatus returns to the starting condition ready to repeat the

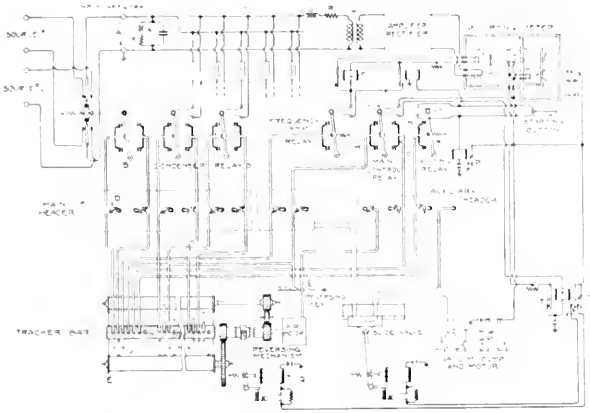


Fig. 3.—Arrangement of Pneumatic and Electrical Apparatus

operation. This is accomplished by means of pneumatic apparatus operating in conjunction with a photographic recording device.

The pneumatic arrangement is a modification of a piano player mechanism in which a paper roll of standard dimensions is used. By proper perforation of the roll special pneumatic relays are operated in proper sequence to switch the condensers of the tuned circuit, flash frequency lines on the record, stop the mechanism after a record has been completed, rewind the piano roll, and perform other functions necessary to leave the analyzer in the starting condition. Electrical relays for switching the tuning condensers were not found practicable on account of the disturbances induced into the analyzer circuit.

The photographic recording apparatus consists of the camera motor for moving the sensitized record paper at a constant rate proper arrangement of lenses and lamps for illuminating the mirror

galvanometer and tracing the scale and frequency lines, and suitable baths for developing and fixing the record. The record is drawn through the mechanism by means of two motor-driven rubber rollers, which also serve to remove excess solution.

The development of the pneumatic switching apparatus was carried out with a view to making use of as many standard piano player parts as possible. However, it was found necessary to make some

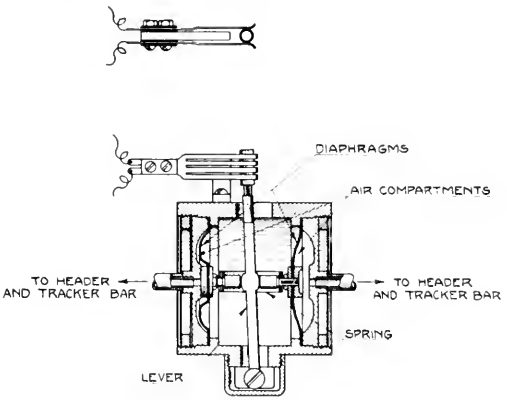


Fig. 4 —Pneumatic Relay

modifications in method and apparatus; in particular a new pneumatic motor element (air relay) for switching the condensers at the requisite speed had to be developed.

Fig. 3 is a schematic drawing showing the principal features of the analyzer. In this drawing the vacuum pump is shown driven by an electric motor, and connected by means of pipes to the auxiliary and main headers and relays. This pump maintains in the headers an absolute pressure of about 4 or 5 lb. per square inch. The player piano roll E operates the entire mechanism by passing over the tracker bar in the usual manner. The air motor and tracker bar equipment are substantially as supplied by the manufacturers except that the reversing mechanism is arranged to be operated electrically instead of by hand.

The essential features of the air relay which was developed for this analyzer may be better understood by reference to Fig. 4. A cylindrical casting is arranged to mount two flexible diaphragms and

two end plates in such a way as to form at each end of the cylinder, compartments, one side of each of which is a diaphragm. When assembled the two diaphragms face each other and are connected together by a circular spring made of steel strip. In use the two end compartments are partially evacuated thus causing the diaphragms to pull apart, straining the spring. When distended the diaphragms lie against the inner faces of the end plates which are shaped as shown. Obviously if air be allowed to enter either of the compartments the diaphragm belonging thereto will be pulled toward the other diaphragm by the spring. Passing through the circular spring is a lever pivoted at one end and carrying on the other end an insulated metallic sleeve. This lever is not attached in any way to either diaphragm and will of itself remain in position where last placed. Switch points are mounted in such a way that the sleeve may be forced in or out between them by the action of the diaphragms. This relay has proved very satisfactory in service and is particularly fast in its operation.

Connections between the tracker bar, main header, and the pneumatic relays are made by means of rubber tubing. As shown in Fig. 3 each of these relays requires two rubber tubes leading to the main header and two from the header to the tracker bar. These tubes are connected to the header by means of stop cocks *D* so connected that the direct passage of air from tracker bar to relay is practically unobstructed but the passage leading from the junction to the header may be made as small as desired by turning the finger valve. As adjusted, the opening to the header is small compared to the size of the tubes so that if air be permitted to enter one of the tube lines (as at the tracker bar), the diaphragm of the relay associated therewith is immediately released. When the tube is closed again, the entrapped air is soon removed through the small opening leading to the header thus restoring the diaphragm to its original position. The relay lever, however, does not follow the diaphragm.

This arrangement possesses the advantage that small openings only are necessary in the player piano roll, and that the opening which connects a condenser into the circuit is not in line on the roll with the opening which disconnects this condenser. Also at the beginning of an analysis by suitable perforations in the roll all air relays can be set simultaneously in the off position (condensers disconnected), thus making sure of the initial conditions. The apparatus is so designed that all the openings causing condenser circuits to close are on one side of the roll and those causing them to open are on the other side.

As before mentioned the tuned circuit is made up of inductance L, having some resistance R, and a bank of condensers designated by C. The function of the "Condenser Relays" is to connect into the tuned circuit any one or any combination of the 25 fixed condensers, thus tuning the circuit in small steps over a wide range of frequencies. The input is fed into this circuit as shown at A, and the degree of resonance, that is the response of the circuit at any

Fig. 5—View of Analyzer Ready for Use

a—Input and tuned circuits
b—Amplifier-rectifier
c—Camera motor
d—Recording meter
e—Control box
f—Starting button
g—Reversing key

particular frequency of tuning, is measured by means of the small transformer T, the amplifier-rectifier and the recording meter.

In addition to operating the tuned circuit a few of the air relays are used to operate the control circuits, mark frequency lines on the chart, etc., uses which required slight modification as indicated schematically in Fig. 3. In two of these control relays only one diaphragm is used, and the switch lever and diaphragm are fastened together by means of a flexible link. It has already been noted

that the analyzer is equipped to trace two curves simultaneously on a single record. This is accomplished by means of air relay *B* which is so arranged as to connect two sources of input alternately to the analyzer. These input connections are alternated rapidly and are effected by appropriate punching of the roll.

The above covers the essential features of the analyzer but there

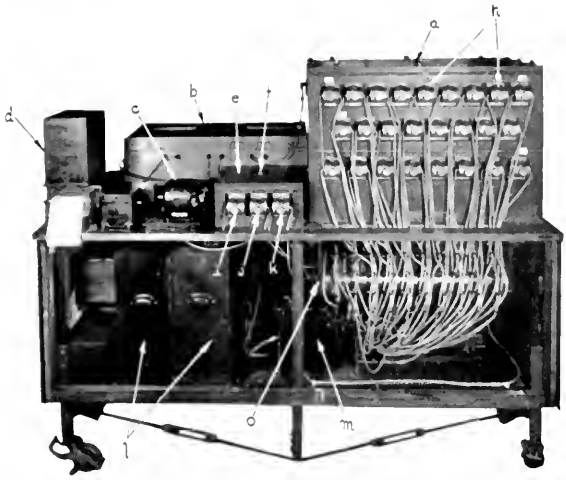


Fig. 6—View of Analyzer with Relay Side Uncovered

a—Input and tuned circuits	*h*—Condenser relays
b—Amplifier-rectifier	*i*—Holding relay
c—Camera motor	*j*—Frequency lamp relay
d—Recording meter	*k*—Main control relay
e—Control box	*l*—Plate battery
f—Starting button	*m*—Air motor
o—Main header	

remain a few details having to do with assembly, control, etc., that may be of interest.

Fig. 5 shows the analyzer as completed and ready to operate. The apparatus is assembled on a two-deck, structural-steel table equipped with castors for convenience in handling. Much of the equipment is inclosed for protection against moisture and dust. The recording meter, camera motor, amplifier-rectifier, control relays, and input and tuned circuits are placed on the top. Below are

mounted the batteries, the vacuum pump with its motor, and the tracker bar with paper roll mechanism. The arrangement is clearly shown in Figs. 6 and 7 taken with the protecting panels removed. In Fig. 6, *a* is a moisture-proof box containing the input and tuned circuits. The inductance coil is placed in the center of the upper half of this box together with the switch for connecting the windings

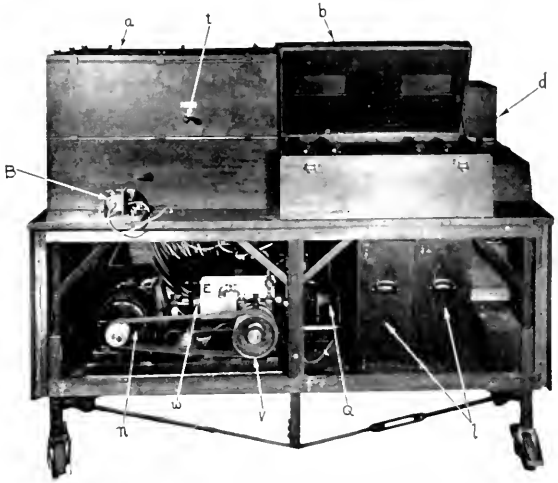


Fig. 7—View of Analyzer with Pump Side Uncovered

a—Input and tuned circuits
B—Amplifier-rectifier
c—Source-alternating relay
d—Recording meter
E—Paper roll
l—Plate battery
n—Pump motor
Q—Reversing solenoid
t—Series-parallel coil switch
v—Vacuum pump
w—Tracker bar

in series or parallel. The smaller capacitances are of mica and are arranged around the coil and switch assembly in such a way that they may be connected with a minimum length of lead to the air relays which are located on one side of the box. The larger capacitance units are made up of paper condensers and are placed in the lower half of the box. The metal lined box *b* contains the amplifier-rectifier, and at *d* is shown the recording meter. Box *e* contains the control circuits with the necessary relays. The method of mounting the air relays, main vacuum header (attached to underside of

table top), air motor, etc., is also clearly shown in this figure. Each air relay is equipped with two rubber tubes leading to adjustable cocks on the header which in turn are connected to the tracker bar. The three-control relays are also shown in Fig. 6. The vacuum pump is shown at v in Fig. 7. The piano roll E moves over the tracker bar w and is reversed by means of solenoid Q. In boxes l are placed the plate batteries for the amplifier-rectifier.

The control apparatus by means of which the analyzer becomes practically an automatic machine will now be described. Referring again to Fig. 3 it will be seen that there is provided an auxiliary header and an electrically operated slide valve. The functions of these devices will be discussed presently.

The machine is started by pressing the starting button which should be kept closed for a few seconds while normal vacuum is being established in the headers. The air motor then starts and the paper roll E begins to travel across the tracker bar. Perforations in the roll are so made that when the roll is in its initial position an opening allows air to enter chamber N of the holding relay. As soon as the paper starts, however, this opening is closed, chamber N is exhausted, and contacts K close. This short-circuits the starting button which the operator may now release, and the machine is in full operation. It will be noted that the closing of the contacts of the starting button or contacts K puts into operation motors which drive the vacuum pump and the camera apparatus. Simultaneously recording meter lamp H and scale-line lamp I are lighted. The latter illuminates the record through small holes in an opaque scale strip thus marking horizontal lines due to the motion of the record.

As the roll E traverses the tracker bar, appropriate perforations control the condenser relays so as to switch the proper condensers into and out of the tuned circuit. Proper perforations also control the frequency lamp relay which flashes frequency lines on the record by means of Lamp G. Relay F is inserted in order to make the flash of short duration.

The tracker bar-paper-roll apparatus was received as a unit from the manufacturer and was installed after making modification in the reversing mechanism as mentioned above. This was done in the interest of automatic control. The paper roll is kept in its proper course over the tracker bar by means of an automatic adjusting device such as used in practically all high grade player pianos.

As the paper progresses over the tracker bar a point is finally reached where the last condenser connections are made and it becomes necessary to rewind the roll and to restore the entire mechanism

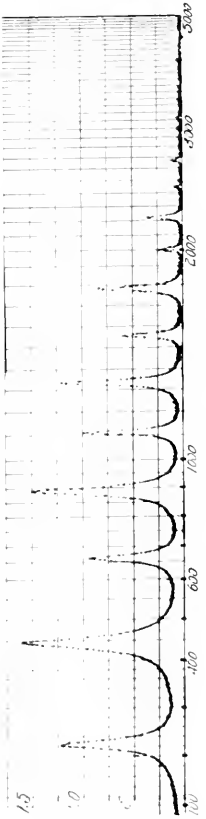


Fig. 8—Output of Carbon Button Driven at an Excessive Amplitude

to its starting condition. This is accomplished by means of a perforation at the end of the record which admits air to chamber M of the main control relay, thus closing contacts J. Relay S then operates since its circuit to ground is completed through contacts P. Operation of relay S opens contacts T thus disconnecting lamps H and I, and closes contacts U and V. It will be seen that the closing of contacts, U operates the reversing mechanism, and rewinding of the roll begins immediately. The closing of contacts V operates the slide valve thus releasing the vacuum on the main header, allowing the roll to be rewound with minimum mechanical drag.

It may be noted that means are also provided for rewinding the roll from any point in its forward travel by admitting air manually at the reversing key. This will cause the main control relay to operate so that rewinding will begin. Vacuum is kept on the auxiliary header during the rewind so that control of the analyzer may be maintained to the end of the operating cycle.

When the paper has been completely rewound perforations allow air to enter simultaneously chamber O of the main control relay and chamber N of the holding relay. This action opens contacts J and K, thus bringing the entire mechanism to rest in its initial starting condition.

Applications

To show the variety of problems in which the analyzer is a useful means of investigation, a few illustrative records have been made and will be discussed. These records were taken in each case to illustrate the use of the analyzer and are not parts of investigations to which they are related. They cannot, therefore, be taken as representative of the performance of the apparatus tested.

One of the uses of the analyzer has been in the study of the performance of microphone buttons. Fig. 8, for example, illustrates the character of the distortion in a button when driven at an excessive amplitude. The button was mounted so that its movable electrode could be driven at a single frequency by a very heavy reed at its natural frequency so that the motion was very nearly sinusoidal. The frequency of the motion was a little less than 450 cycles corresponding to the second peak on the record. The amplitude of motion was 0.004 centimeters or 0.0004 inches which is of course much greater than normally obtains in a transmitter. The circuit consisted simply of the button and a battery in series with the analyzer so that the record is an analysis of the current fluctuations in the button. The record shows two series of frequencies generated

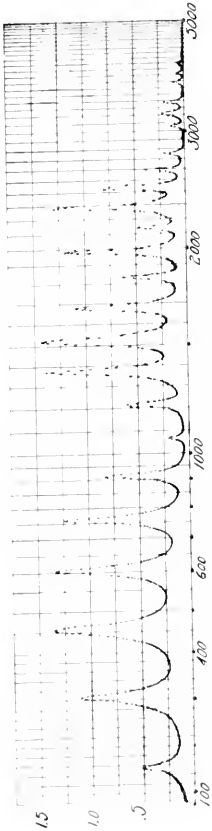


Fig. 9—Noise in Room as Picked up by Condenser Transmitter

Fig. 10—Noise in Room as Picked up by Telephone Receiver Used as a Transmitter

by the button; a primary series having for its fundamental the driving frequency, 450 cycles, and a subsidiary series, having for its fundamental half the driving frequency or 225 cycles. The even harmonic components of the secondary series coincide, of course, with the frequencies of the primary series. The primary series can be accounted for by the fact that with such large amplitudes the changes in resistance are not a linear function of the amplitude of

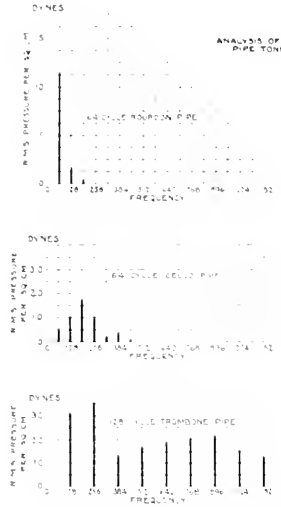


Fig. 11—Analysis of Organ Pipe Tones

motion. The subsidiary series is due to the non-symmetrical effect of the inertia of the carbon grains in vibration, the motion being so violent that some of the grains are thrown free from their contacts. For small amplitudes such as those ordinarily encountered in a transmitter, a record would show only 450 cycles, the other frequencies occurring in negligible amount; for intermediate amplitudes the primary series only occurs.

The analyzer has been used in connection with the study of sustained sounds and of the performance of acoustical apparatus. Fig. 9 is a record of the noise in a room originating from a buzzer as

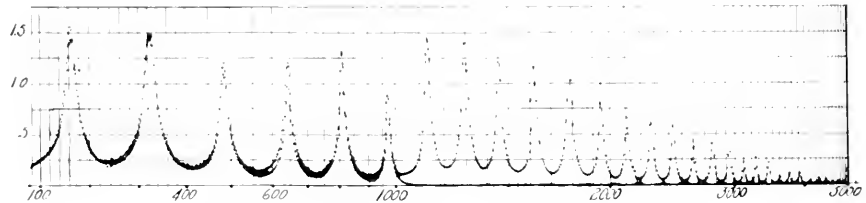


Fig. 12—Record Showing Action of Low Pass Filter

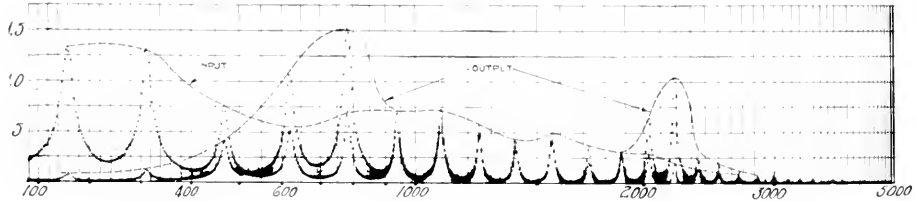


Fig. 13—Records of Electrical Input and Acoustic Output of a Common Type of Loud Speaking Receiver

picked up by a condenser transmitter.[3] The reverberation in the room probably had a large effect on the character of this record. With such a source of frequency the analyzer may be used to study the acoustics of rooms. Fig. 10 is a record of the same noise as in Fig. 9 but as picked up by a common type of telephone receiver placed in the same position as the condenser transmitter. A comparison of Figs. 9 and 10 will show the inadaptability of such a receiver for use as a transmitter. The receiver, owing to the resonance of its diaphragm, is seen to be relatively sensitive in the region of 600 to 800 cycles and insensitive at most other frequencies. When this instrument is placed against the ear, as when used as a receiver, the diaphragm resonance is damped so as to give more nearly uniform response.

By means of the calibration of the condenser transmitter and its amplifier, it is possible to make an analysis of the absolute intensity of a sustained sound in the air. This method has been used to study the frequency characteristics of musical instruments. Fig. 11 shows the analyses of three low-frequency organ pipes. These are plots of r. m. s. pressure change in the sound wave as obtained from the analyzer records. Each vertical line corresponds to a peak on the original record. The upper chart shows the almost pure tone given by a 64-cycle Bourdon pipe. In the case of the cello pipe, also having a fundamental of 64 cycles, the third harmonic is seen to be more prominent than the fundamental or second harmonic. The third chart is for a 128-cycle trombone pipe which was found to be rich in harmonics. The pressure in the single components of the cello and trombone pipes is less than in the case of the Bourdon pipe, and a larger scale of ordinates is therefore used.

To illustrate the use of the attachment which permits the making of two simultaneous analyses, a few double records will be presented. An electric wave filter which has been used in the study of telephone quality was connected to the buzzer source whose output is shown in Fig. 2. Simultaneous analyses of the current delivered to and transmitted through the filter are shown in Fig. 12. This filter is designed to pass all frequencies below 1000 cycles and to suppress all others. The input is represented by a more or less continuous series of peaks along the entire length of the record. The peaks corresponding to the output coincide rather closely with the input

[3] "A Condenser Transmitter as a Uniformly Sensitive Instrument for the Absolute Measurement of Sound Intensity," E. C. Wente, *Physical Review*, July 1917.

"The Sensitivity and Precision of the Electrostatic Transmitter for Measuring Sound Intensities," E. C. Wente, *Physical Review*, May 1922.

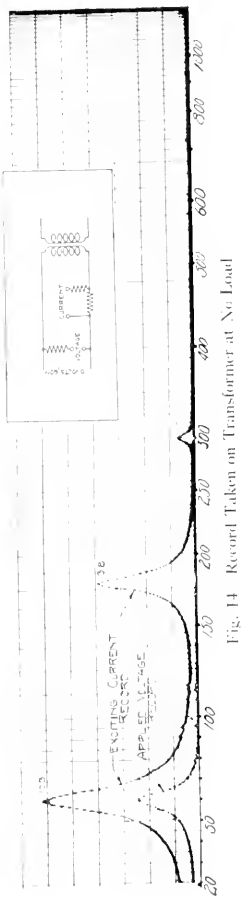


Fig. 14—Record Taken on Transformer at No Load

Fig. 15—Record Taken on Transformer Under Load

peaks for all frequencies below 1000 cycles and are not detectable for the higher frequencies.

Fig. 13 is a double record showing the analysis of the wave from a buzzer as fed into a common type of loud speaking receiver and the acoustic output as picked up by a condenser transmitter placed in front of it at a distance of about 15 inches. The analysis of the input current wave to the loud speaker is shown by the comparatively continuously decreasing series of peaks. The acoustic output is represented by the series having maxima in the neighborhood of 800 cycles and 2200 cycles. This record cannot be taken as an adequate analysis of this loud speaker because of probable reverberation effects in the room.

The analyzer has thus far not been used in the study of power problems. A few illustrative records have been taken, however, on transformers and generators and will be shown as suggestive of the use of this method of attack in such problems.

Fig. 14 is a double record showing applied voltage and exciting current of a small 110-volt, 60-cycle transformer operating at normal voltage and frequency under the no-load condition. The presence of the well known third and fifth harmonics in the exciting current is clearly shown. Because of the rise in the calibration curve of the analyzer at the low end of the lower frequency range, a scale of ordinates is not shown on this record. Instead, the values of the analyzer current at each frequency are noted on the record. The circuit used in making this record is drawn on the figure. A computation of the components of the exciting current from the record and constants of the circuit shows that at 60 cycles the current was 175 milliamperes, at 180 cycles, 65 milliamperes and at 300 cycles, 17 milliamperes. The total r. m. s. exciting current was therefore 187 milliamperes.

The operation of this transformer under full load is shown in Fig. 15, where, as before, the primary voltage and current are analyzed. The transformer load consisted of a pure resistance. It will be noted that the third and fifth harmonics have become very small compared with the fundamental. The analyzer currents at each frequency are again noted on the record. In obtaining the analysis of the current it was necessary to further shunt the analyzer. The primary current was 310 milliamperes.

Problems relating to commutation may also be conveniently studied qualitatively and quantitatively by means of the analyzer. The use of an apparatus which will indicate the source and measure the extent of parasitic frequencies is obvious. Information has

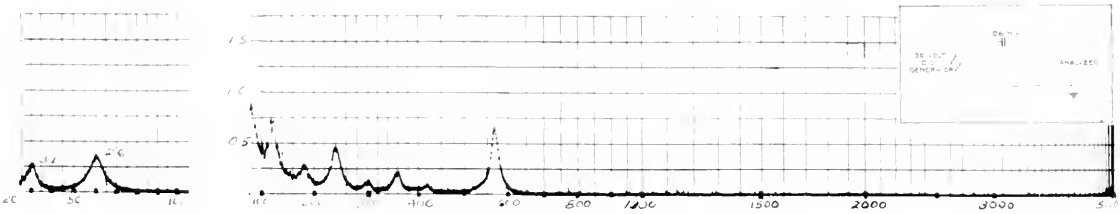


Fig. 16—Record Taken on D. C. Generator at No Load

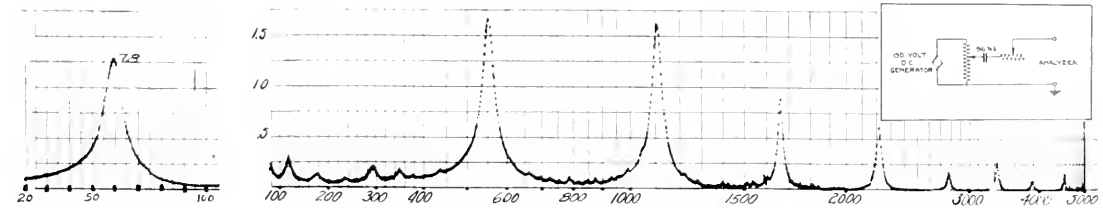


Fig. 17—Record Taken on D. C. Generator Under Load

been obtained on a small machine direct-driven by a $\frac{1}{2}$-h. p., 60-cycle single-phase motor. Data of importance relating to the generator tested are as follows:

Capacity of Generator	$\frac{1}{4}$ kw.
Number of Poles	2
Speed	1725-1800 r. p. m.
Voltage	125
Field	Shunt-connected
Diameter of Commutator	2.75 in.
Number of Commutator Bars	38
Number of Armature Slots	19
Size of Brush	$\frac{3}{8}$ in. square
Yoke	Ring type

Records obtained from this machine when operating under no-load and half-load conditions are shown in Figs. 16 and 17, respectively. The corresponding speeds are approximately 1800 and 1750 r. p. m. In order to show what frequencies the machine gives out over the entire range 20 to 5000 cycles each figure is made up of two parts: a portion of a 20-1250 record and a complete record over the range 80-5000 cycles. On each figure is drawn the circuit connecting the d-c. generator to the analyzer. It will be noted that a large condenser is inserted to prevent the passage of heavy direct current through the analyzer.

The consideration of these records leads to the conclusion that there are at least three independent major causes of alternating voltage operating in this d-c. machine. The fundamental frequencies due to these causes are 30, 60 and 570 cycles. It will be noted that the 30-cycle peak occurs only on the no-load record under which condition the average speed is practically 30 revolutions per second. Sixty cycles and a series of its harmonic overtones are seen to be present under both conditions of load. Under load the 60 cycles is augmented whereas its harmonics are reduced. No harmonic overtones of 30 cycles except such as might coincide with the harmonics of 60 cycles are found in either case. This indicates the existence of independent causes of the 30 and 60-cycle frequencies, that the 30-cycle cause produces an almost sinusoidal voltage, and that the 60-cycle cause under no load produces an irregular wave which becomes smoother as the machine is loaded.

The no-load record, Fig. 16, shows 570 cycles with no harmonics while the load record, Fig. 17, shows 570 cycles with a complete series of harmonics. This indicates that at no load the cause of 570 cycles

feeds a relatively smooth wave to the line while under load this cause feeds an irregular wave to the line. The fact that 1140 cycles is about as strong as the fundamental and that its harmonics are stronger than alternate ones which are overtones of 570 only, suggests the likelihood of a fourth cause having a frequency of 1140 cycles. Small irregularities at frequencies other than those already mentioned occur in the record. These are more prominent under load than at no load and indicate the presence of small, more or less irregular pulses, which increase with load. All of the above frequencies may be accounted for by a consideration of the construction and operating condition of the machine.

The generator was driven by a single-phase, 4-pole, 60-cycle motor which may give rise to torque fluctuations once per revolution, or 30 times per second. Under no load this may produce considerable corresponding fluctuations in speed while under load conditions the generator acts as a damper, eliminating these oscillations.

The 60-cycle peak may be due to any one or some combination of a number of causes, *e. g.*, eccentricity of generator armature, non-uniform winding, non-uniform thickness of mica separators in commutator, high mica between one or more pairs of segments, etc. The records show that for this particualr machine in its present condition (new) at normal speed the 60-cycle voltage developed increases considerably with load indicating strongly that the cause is largely influenced by an IR drop somewhere in the machine. The most likely causes therefore appear to be commutator eccentricity, irregular spacing of the segments, or high mica.

The peak at 570 cycles may be accounted for by cyclic variation of flux entering the armature core as the teeth pass the pole faces. At no load the speed is approximately 1800 r. p. m. The number of teeth being 19, it is obvious that there will be 570 fluctuations of air-gap reluctance per second. Under no-load conditions the record shows a comparatively pure wave form for this cause. This is to be expected because of the comparatively uniform distribution of flux under the pole faces at no load. As the machine is loaded, however, the field is distorted and shifted giving rise to an irregular wave form of voltage which is responsible for at least a part of the large harmonic content shown by the load record.

The presence of 1140-cycle peak which is present only under the load condition may be due to the cyclic variation of voltage produced by the commutator bars leaving the brushes. Inasmuch as the speed is roughly about 29 revolutions per second the frequency with which bars leave brushes is about 1100 cycles. This frequency is present

under the load condition only, thus indicating that it is due to an IR drop at the brush contacts or to an e. m. f. developed in the short-circuited coil with the brush off the magnetic neutral.

The very small irregularities on the record shown particularly between peaks above 550 cycles on the load record are probably due to slight chattering of the brushes.

It is of interest to note that the so called frequency of commutation does not appear in either of the records. For this machine this frequency at no load is approximately 316 cycles per second.

From these records it is possible to determine the r. m. s. value of the alternating voltage at any frequency of interest. This is computed from a knowledge of the circuit constants and analyzer impedance. We thus obtain for the 550-cycle peak (Fig. 17) a value of 0.8 volts and for the 60-cycle peak a value of 1.1 volts.

In general the records taken by means of the analyzer on this commutating machine, confirm quantitatively the well known fact that such machines may give rise to frequencies in the audible range. Consideration of the records indicates that these frequencies may be divided into two classes: First, those pertaining to and controlled by design, and second, those caused and controlled by the physical condition of the machine at any particular time. It is also interesting to note that the driving motor may produce an appreciable effect, particularly under the no-load condition.

Summary

In the above paper there has been given a short statement of the theory and construction of an automatic, recording, electrical frequency analyzer, together with illustrations showing its use and limitations in various fields.

This apparatus has been found very useful in the laboratory in the investigation of many different types of problems chiefly because of the speed with which records can be made and harmonic analyses obtained without computation.

In conclusion the authors wish to express their appreciation to Mr. C. E. Lane and Mr. C. E. Dean, of the Western Electric Company, Inc., for their assistance in the building of this machine and the preparation of this paper

Certain Factors Affecting Telegraph Speed[1]

By H. NYQUIST

Synopsis: This paper considers two fundamental factors entering into the maximum speed of transmission of intelligence by telegraph. These factors are signal shaping and choice of codes. The first is concerned with the best wave shape to be impressed on the transmitting medium so as to permit of greater speed without undue interference either in the circuit under consideration or in those adjacent, while the latter deals with the choice of codes which will permit of transmitting a maximum amount of intelligence with a given number of signal elements.

It is shown that the wave shape depends somewhat on the type of circuit over which intelligence is to be transmitted and that for most cases the optimum wave is neither rectangular nor a half cycle sine wave as is frequently used but a wave of special form produced by sending a simple rectangular wave through a suitable network. The impedances usually associated with telegraph circuits are such as to produce a fair degree of signal shaping when a rectangular voltage wave is impressed.

Consideration of the choice of codes show that while it is desirable to use those involving more than two current values, there are limitations which prevent a large number of current values being used. A table of comparisons shows the relative speed efficiencies of various codes proposed. It is shown that no advantages result from the use of a sine wave for telegraph transmission as proposed by Squier and others[2] and that their arguments are based on erroneous assumptions.

Signal Shaping

SEVERAL different wave shapes will be assumed and comparison will be made between them as to:

1. Excellence of signals delivered at the distant end of the circuit, and

2. Interfering properties of the signals.

Consideration will first be given to the case where direct-current impulses are transmitted over a distortionless line, using a limited range of frequencies. Transmission over radio and carrier circuits will next be considered. It will be shown that these cases are closely related to the preceding one because of the fact that the transmitting medium in the case of either radio or carrier circuits closely approximates a distortionless line. Telegraphy over ordinary land lines

[1] Presented at the Midwinter Convention of the A. I. E. E., Philadelphia, Pa. February 18, 1924, and reprinted from the Journal of the A. I. E. E. Vol. 43, p. 124, 1924.

[2] A. C. Crehore and G. O. Squier. "A Practical Transmitter Using the Sine Wave for Cable Telegraphy; and Measurements with Alternating Currents upon an Atlantic Cable." A. I. E. E. Trans., Vol. XVII, 1900, p. 385.

G. O. Squier. "On An Unbroken Alternating Current for Cable Telegraphy." *Proc. Phys. Soc.*, Vol. XXVII, p. 540.

G. O. Squier. "A Method of Transmitting the Telegraph Alphabet Applicable for Radio, Land Lines, and Submarine Cables." *Franklin Inst., JL.*, Vol. 195, May 1923, p. 633.

employing direct currents will next be considered. This will be followed by a consideration of the more complicated case of transmission over long submarine cables.

It will be shown that the waves produced by sending rectangular signal elements through suitable electrical networks which round them off before they are impressed on the transmitting medium are probably best in most cases. Comparsion will be made between waves shaped by sending rectangular signal elements through suitable networks and waves made up of half cycles of a sine wave, bringing out the inferiority of the latter.

Direct-Current Telegraph Transmission Over a Distortionless Line

Before proceeding with this discussion two terms, which will be used in this paper, and which are considered to be of fundamental importance, will be defined—"signal element" and "line speed." It is usually possible, especially when sending is done mechanically, to divide the time into short intervals of approximately *equal* duration, such that each is characterized by a definite, not necessarily constant, voltage impressed at the sending end. The part of the signal which occupies one such unit of time will be called a "signal element." For example, the letter *a* in ordinary land telegraphy will be said to be made up of five signal elements, the first constituting a dot, the second a space and the next three a dash. The "line speed," as used in this paper, equals the number of signal elements per second divided by two. In ordinary land telegraphy the line speed is equal to the dot frequency when a series of dots separated by unit spaces is transmitted.

The discussion will first be limited to the case of direct-current telegraphy over a distortionless line. This case is the simplest, and in addition the results will aid in understanding the more complex cases. It may aid in obtaining an understanding of this case to assume that the distortionless line is made up simply of series and shunt resistances.

A distortionless line, such as the one which has been assumed, will transmit all frequencies with equal efficiency from zero upward. In considering applying direct-current telegraph to this line, it will be assumed that the telegraph circuit will have assigned to it only a limited range of frequencies from zero upward, the remaining frequency range being assigned to some other uses, such as ordinary telephone and carrier telephone and telegraph. It will also be as-

sumed that the direct-current telegraph circuit is worked at as high a speed as the frequency range assigned to it will permit.

A number of different wave forms which might be employed to make up the telegraph signal elements will next be examined, consideration being given first to the waves which will be received at the distant end when the different wave forms are impressed at the transmitting end and second to the interference which will be produced in the higher range of frequencies which has been assigned to other uses.

Three forms of voltage waves which will be considered are shown in Fig. 1. *A* in that figure shows the simplest form of voltage wave,

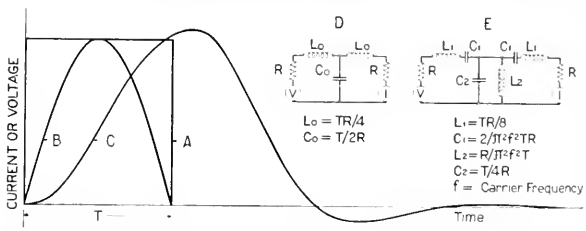


Fig. 1

A—Rectangular Voltage Wave
B—Half Cycle of Sinusoidal Voltage Wave
C—Rectangular Voltage Wave Modified by Being Passed through Network Shown at *D* or *E*.

namely, the rectangular form which is produced by applying a battery for a given interval of time and then substituting a short circuit for it. *C* in the figure is the wave produced by transmitting the rectangular voltage wave *A* through an electrical network which is the one indicated by the letter *D* in the figure. (Other forms of networks might also be selected which would produce similar results.) *B* in the figure is a wave which has the shape of a half cycle of a sine wave. In what follows this wave will be referred to as the "half-cycle sine wave."

In considering the waves which will be received when the above waves are applied at the transmitting end, use will be made of the following general principles, which have been stated by Malcolm,[3] for the case of a submarine cable circuit and discussed for the general case in Appendix A.

[3] H. W. Malcolm. "Theory of the Submarine Telegraph and Telephone Cable." The Electrician Printing & Publishing Co., London, March 1917.

When a telegraph circuit is worked at a line speed as high as will be permitted by the available frequency range, the shape of the received signal will be practically independent of the shape of the transmitted signal, and further, the magnitude of the received signal will be approximately directly proportional to the area included within the impressed voltage wave.

The area included within the impressed voltage wave being of principal importance so far as the wave received at the distant end is concerned, the areas under the three voltage waves shown in Fig. 1 will next be examined. The areas under waves A and C will be found to be substantially equal while the area under the wave B is only about 0.6 as great. Consequently, it should be expected that waves A and C will be about equally good from the standpoint of the received signals, while wave B will be poorer, producing received signals only about 0.6 as great in magnitude. If the maximum voltage (or power) impressed at the sending end is limited to some given value, the rectangular wave is seen to be the optimum, since this wave has the maximum area. While the area shown under curve C is approximately equal to that under the rectangular wave, the effect produced when a number of signal elements of the same polarity and magnitude are sent in succession is such that the maximum voltage transmitted will exceed slightly the corresponding voltage for the case of the unmodified rectangular wave due to overlapping of adjacent signal elements.

The above comparison of the three waves of Fig. 1 from the standpoint of received signals holds not only for signal elements, but also for complex waves comprising a number of elements. Since for the speeds under consideration the received currents for different shapes of signals applied at the sending end are substantially of the same form, differing, at most, in magnitude, it follows from the principle of superposition that any complex signal, whether built up of elements of one shape or another at the sending end, will produce substantially the same wave form at the receiving end, the differences in the shapes of the elements at the sending end producing differences principally in magnitude of the received waves.

Consideration will next be given to the relative interference which the different wave forms of Fig. 1 will produce in the frequency range assigned to other circuits. Since interference into other circuits results from having the telegraph signal elements contain frequencies which spread into the ranges assigned to other circuits, it is evident that the wave will be the best from the standpoint of interference which contains the least amount of these outside frequencies. By

making use of a method which is discussed in Appendix C, the frequency components of the three waves illustrated in Fig. 1 have been computed and are shown in Fig. 2. The frequency marked 1 2 T in the drawing equals the line speed. T in this connection has the same value as in Fig. 1. The letters in this figure refer to the corre-

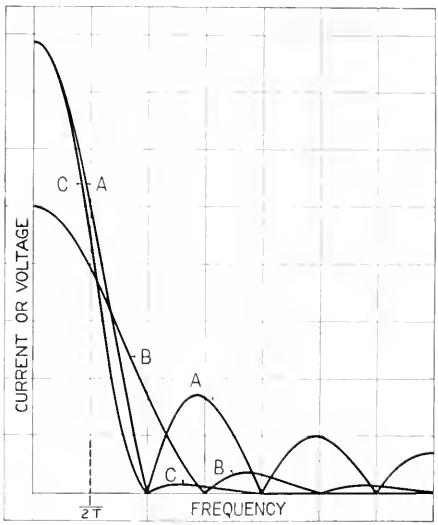


Fig. 2

A Frequency Components of a Single Dot, Rectangular Wave
B Frequency Components of a Single Half Cycle of a Sine Wave
C Frequency Components of a Single Dot, Rectangular Wave Passed through Network Shown in Fig. 1

sponding waves in Fig. 1, A being the components of an isolated rectangular wave, B the corresponding components for the half-cycle sine wave, and C those for the rectangular wave after it has been transmitted through the network D in Fig. 1. It is seen from Fig. 2 that the rectangular wave form A contains the greatest amount of currents of higher frequencies and is, therefore, the poorest from the standpoint of interference. The half-cycle sine wave contains less of these higher frequencies although, as will be seen, the high-frequency components are far from negligible. The wave C is the

best from the standpoint of interference, since it contains the least amount of these higher frequencies.

From the preceding it is concluded that for the case under consideration, the wave form C in Fig. 1 produced by sending a rectangular shaped signal element through a suitable network is the most suitable. This wave form is almost the optimum from the standpoint of the received signals while from the standpoint of interference into other circuits it leaves little to be desired.

Carrier and Radio

The results for the distortionless line are particularly applicable to the cases of radio and carrier telegraphy because in these cases we have a transmitting medium which is substantially distortionless. We may again make use of Fig. 1 to illustrate three possible voltages, it being understood that these curves represent the envelope or outline of the transmitted currents which are in reality of a frequency considerably higher than the signaling frequency. If now we limit consideration to the case where the carrier frequency is located in the middle of the transmitted frequency band, then, this case becomes very similar to the direct-current case and what has been said about the received wave shape being independent of the transmitted one and its magnitude being directly proportional to the area under the transmitted voltage curve still holds. One important difference is that, whereas in the direct-current case the network shown at *D*, Fig. 1, is used in the alternating-current case having the carrier located in the middle of the free transmitted range, the network shown at *E*, Fig. 1, is used. A further difference is that in the case of radio where very high frequencies are involved, it may not be practicable to construct the required networks. In that case, however, it is practicable to produce the corresponding direct-current wave and utilize it to modulate the radio wave.

What was said about interference from the circuit in question into other circuits in the direct-current case above also holds for the case of radio and carrier with the difference that whereas Fig. 2 shows a band of frequencies extending from zero up, the corresponding curve in the case of radio and carrier consists of two such bands. The complete curve for radio and carrier is substantially symmetrical with respect to the ordinate corresponding to the carrier frequency, and the right-hand portion is similar to the curve shown in Fig. 2. It will be obvious that the rectangular wave and the half-cycle sine wave are both objectionable, as voltage waves to be applied to the

transmitting medium, because they contain frequency components which may easily extend into the range allotted to neighboring carrier bands. For this reason it is customary in carrier telegraph practise to make use of a transmitting filter to cut off these interfering frequencies. The voltage impressed on this filter is substantially rectangular in outline but after passing the filter it has a shape which is approximately similar to curve C in Fig. 1, and which, therefore, produces less interference than a half-cycle sine wave.

Land Lines

The case of land lines is somewhat different from the case discussed previously because it is not economically desirable to utilize the full frequency range available. In other words, the great expenditure for terminal apparatus that may be proper in the case of submarine cables and long distance radio circuits is not warranted. In land circuits the highest frequencies transmitted are considerably greater than the required line speed. When this is the case, it is usually possible and desirable to make use of the available range to increase the steepness of the received wave. A steep wave front results in prompt operation of the receiving relay and this in turn results in minimum distortion. If a half-cycle sine wave were to be employed instead of the usual rectangular wave or if a network were to be employed which were to round off the wave to the extent indicated in Fig. 1, the received wave would necessarily lose a great part of its steepness and as a consequence the response of the receiving relay would be less positive and the signals would be distorted. It will, of course, be understood that by means of suitably proportioned networks the wave can be rounded just enough to meet the interference requirement, still retaining sufficient steepness to insure prompt operation of the receiving relay. Therefore, rounding by means of networks is preferable.

If it should be desirable and practicable to utilize the frequency range to its fullest, what has been said above about a distortionless line holds without any substantial modification and it would, in that case also, be more advantageous to use a wave rounded by means of suitable networks than to impress on the line a wave of the half-cycle sine form.

Submarine Cables

In the case of submarine-cable telegraphy, there is a limitation on voltage which has not been emphasized in the simple direct-current case discussed above. The voltage which may be impressed on the

cable is limited to a definite value. Moreover, for certain reasons, the cable has an impedance associated with it at the sending end which may make the voltage on the cable differ from the voltage applied to the sending-end apparatus. Inasmuch as the limitation in this case is voltage limitation at the cable, the ideal wave is one which applies a rectangular wave to the cable rather than to the apparatus, because it insures that the area under the curve should be the maximum consistent with the imposed limitations. It would be possible to make the transmitting-end impedance approximately proportional to the cable impedance throughout most of the important range. This would insure that the wave applied to the cable would have approximately the same shape as the wave applied to the apparatus. It would probably be desirable for practical reasons to make this impedance infinite for direct current.

In connection with the submarine cable a special kind of interference is particularly important, namely, that due to imperfect duplex balance. For a given degree of unbalance, the interference due to this source may be reduced by putting networks either in the path of the outgoing current or in the path of the incoming current. These facts, together with the frequency distributions deduced above for each of the several impressed waves as exhibited in Fig. 2, make it apparent that the beneficial reaction on the effect of duplex unbalance, which can be obtained by the use of a half-cycle sine wave instead of a rectangular wave, can be obtained more effectively by the use of a simple network, either in the path of the outgoing or in the path of the incoming currents. Either of these locations is equally effective in reducing interferences from duplex unbalance, but the location of the network in the path of the outgoing current has the advantage that it decreases the interference into other circuits, whereas the location in the path of the incoming current has the effect of reducing the interference from other circuits.

Before leaving the matter of submarine telegraphy, it may be well to point out that it is common in practise to shorten the period during which the battery is applied so as to make it less than the total period allotted to the signal element in question. For instance, if it is desired to transmit an *e* the battery may be applied for, say, 75 per cent. of the time allotted to that *e* and during the remaining 25 per cent. the circuit is grounded. The resulting voltage is shown in Fig. 3F. From the foregoing, it is concluded that this method is less advantageous than the application of the voltage for the whole period, because while the shape of the received signal is substantially the same in the two cases, the magnitude, being proportional to the area under

the voltage curve, will be less. A cursory examination of the literature does not disclose that anything has been published on the experimental side either to confirm or to oppose this result.

Choice of Codes

A formula will first be derived by means of which the speed of transmitting intelligence, using codes employing different numbers of current values, can be compared for a given line speed, *i.e.*, rate of sending of signal elements. Using this formula, it will then be shown that if the line speed can be kept constant and the number of current values increased, the rate of transmission of intelligence can be materially increased.

Comparison will then be made between the theoretical possibilities indicated by the formula and the results obtained by various codes in common use, including the Continental and American Morse codes as applied to land lines, radio and carrier circuits, and the Continental Morse code as applied to submarine cables. It will be shown that the Continental and American Morse codes applied to circuits using two current values are materially slower than the code which it is theoretically possible to obtain because of the fact that these codes are arranged so as to be readily deciphered by the ear. On the other hand, the Continental Morse code, as applied to submarine cables, or other circuits where three current values are employed, will be shown to produce results substantially on par with the ideal. Taking the above factors into account, it will be shown that if a given telegraph circuit using Continental Morse code with two current values were rearranged so as to make possible the use of a code employing three current values, it would be possible to transmit over the rearranged circuit about 2.2 times as much intelligence with a given number of signal elements.

It will then be pointed out why it is not feasible on all telegraph circuits to replace the codes employing two current values with others employing more than two current values, so as to increase the rate of transmitting intelligence. The circuits, for which the possibilities of thus securing increases in speed appear greatest, are pointed out, as well as those for which the possibilities appear least.

Theoretical Possibilities Using Codes with Different Numbers of Current Values

The speed at which intelligence can be transmitted over a telegraph circuit with a given line speed, *i.e.*, a given rate of sending of signal

elements, may be determined approximately by the following formula, the derivation of which is given in Appendix B.

$$W = K \log m$$

Where W is the speed of transmission of intelligence,
m is the number of current values,
and, K is a constant.

By the speed of transmission of intelligence is meant the number of characters, representing different letters, figures, etc., which can be transmitted in a given length of time assuming that the circuit transmits a given number of signal elements per unit time.

Substituting numerical values in this formula gives the following table which indicates the possibilities of speeding up the transmission of intelligence by increasing the number of current values.

Number of Current Values Employed	Relative Amount of Intelligence which can be Transmitted with a Given Number of Signal Elements
2	100
3	158
4	200
5	230
8	300
16	400

This table indicates that there is considerable advantage to be secured in going to more than two current values where the circuits are such as to permit it and where the line speed is not lowered as a result. The limitations will be outlined below. It should also be noted that whereas there is considerable advantage in a moderate increase in the number of current values, there is little advantage in going to a large number.

Codes Now in Common Use—Comparison with Ideal

In the case of printer codes, the theoretical results derived correspond closely to practise, as will be obvious from the method of deriving the formula.

In order to compare the theoretical possibilities indicated by the formula with the results which are obtained when non-printer codes are constructed, several codes were assumed, and for each one the number of signal elements required to produce an average letter

was deduced. The method of doing this is set forth in Appendix D. This work resulted in the following table:

	Signal Elements per Letter	Relative Number of Letters for a Given Number of Signal Elements
American Morse (two current values)	8.26	74
Continental Morse (two current values)	8.45	73
Ideal (two current values)	6.14	100
Continental Morse (three current values)	3.77	163
Ideal (three current values)	3.63	169

The column in the above table headed "Relative Number of Letters for a Given Number of Signal Elements" makes possible direct comparison with the results predicted from the formula as given in the table which preceded. It will be noted that the ideal three-current-value code gives an increase in the number of letters for a given number of signal elements as compared with the ideal two-current-value code which is in fair agreement with the theoretical ratio of 1.58:1. It will also be noted that the Continental three-current-value code which is actually in use in the case of submarine cables appears to come quite close to the ideal. In the case of the Continental and American Morse codes, however, where only two current values are used, the results fall short of the ideal, the ratio between the results actually obtained and the ideal being approximately 1.4:1. The reason for this is that a certain proportion of the possible speed is sacrificed in order to make it possible to read the signals by means of a sounder instead of recording them. For instance, the dash has been assumed to be approximately three times as long as the dot. If the signals were mechanically formed at the sending end and recorded at the receiving end, it would be possible to make use of markings 1, 2, 3, etc., signal elements long, as well as corresponding spacings. The ideal codes were so constructed.

It will be seen that the figures deduced for the Continental Morse and the American Morse are substantially identical for two current values. This result probably does not correspond with practise; it is thought that the difference in speed between these two codes is considerably greater, say on the order of 10 or 15 per cent. in favor of the American Morse. The discrepancy is due partly to the fact that no account has been taken of figures and punctuation marks in the present computations and partly to the fact that the assumptions as to relative lengths of space is not strictly in accordance with practise.

From the foregoing, it is seen that there is a two-fold gain in changing from the two-current-value American or Continental Morse codes to the three-current-value Continental code. In the first

place, there is a theoretical increase in the ratio of 1.6:1 which accompanies the change from the two-current-value to the three-current-value code. In the second place, there is an incidental increase in the ratio of 1.1:1, due to the fact that the present two-current-value codes are longer than would be necessary, if receiving were done by means other than the ear. The total increase in going from the two-current-value Continental or American Morse codes to the three-current-value Continental code is, therefore, in the ratio of 1.6×1.1:1 or 2.2:1, provided the line speed is the same. In this connection it should be noted that in the case of the American Morse, the ratio is probably somewhat less than this for the reasons pointed out above.

Limitations in Applying Codes with More than Two Current Values

Certain inherent limitations which have to do with how much the number of current values can be advantageously increased are as follows:

1. Fluctuations in transmission efficiency of the circuit,
2. Interference,
3. Limitations on the power or voltage which it is permissible to employ.

In addition it may be stated that, in general, whenever more than two current values are employed it is necessary to make the sending and receiving means more complicated and expensive. There may be nothing to gain, therefore, in using codes other than those made up of two current values where the telegraph circuits are cheap.

Considering now the features which limit the number of current values which can be employed, it is believed that the importance of the first factor will be obvious. If the line is subject to fluctuations so that the stronger currents at certain times become less in magnitude than the weaker currents at other times, it will be impossible to discriminate between the different current strengths making up the code, particularly if the fluctuations are rapid.

In connection with interfering currents, it is evident that these may be of such polarity as to add to or subtract from the signaling currents and it is consequently necessary to separate the various current values employed sufficiently so that one current value with the interference added may be distinguished from the next larger current value with the interference subtracted.

The spacing between the current values being determined by the interference and fluctuations in transmission efficiency, it will be

seen that the maximum number of current values which can be employed is determined by the maximum power which it is permissible to use.

In the case of land line telegraph circuits operated with direct currents, it is well known that quadruplex circuits are much more seriously affected by fluctuations and interference than are circuits employing only two current values. (A quadruplex telegraph circuit employs four current values for transmission in one direction.) In general, it may be said that the possibilities of improving ordinary direct-current operated telegraph circuits in this manner do not appear particularly promising.

In the case of wireless transmission over great distances all three of the above factors are important in limiting the number of current values which can be effectively employed. In the first place, as is well known, large variations take place in the efficiency of the transmitting medium so that the received signals vary considerably in magnitude from time to time. Secondly, the interference, at least at certain seasons, is great enough to make it difficult to distinguish between the current values even when the usual method which employs only two current values is employed. Thirdly, the received power is limited because of the great attenuation suffered by the wireless waves.

In the case of carrier transmission, it may be that there will be a field for the use of more than two current values. The relative cheapness of the line circuits, however, will tend to limit the amount by which it will be economical to increase the cost and complexity of the receiving apparatus. Moreover, it should be borne in mind that no allowance has been made for the effect on the line speed of increasing the number of current values, this being considered outside the scope of the present paper.

Changing an existing network of telegraph circuits so as to employ a code with three instead of two current values would require new types of telegraph repeaters as well as new sending and receiving apparatus, and new operating methods. It is considered to be outside of the scope of this paper to go into a discussion of the details of this matter.

"Sine Wave" Systems

Considerable interest and discussion has been created by suggestions which have been made to use so-called "sine wave" systems of telegraphy. In view of this, a brief discussion of these systems is given below.

A brief analysis of what are the fundamental features of these systems will be given and, based on the results which have been developed in the preceding discussion, comparison will be made of these systems with systems based on other principles. A particular effort will be made to clear up what appears to be fundamentally incorrect assumptions which underlie the arguments which have been advanced in favor of these "sine wave" systems.

Crehore-Squier System. The use of a sine wave envelope to improve the characteristics of telegraph signals was advocated by Crehore and Squier.[4] The words "United States" formed by means of a wave of this type are shown in Fig. 3*d*. The code employed is the same as the ordinary Continental Morse, the only difference being that the signal elements consist of half-cycle sine waves.

In what has preceded, it has been shown that a half-cycle sine wave has a smaller area than a rectangular wave rounded off by passing through an electrical network and, consequently, the sine wave is inferior to the latter from the standpoint of the received signals. From the standpoint of interference into other circuits, it has also been pointed out that the half-cycle sine waves contain more high-frequency components than properly rounded off rectangular waves. Consequently more interference into other circuits will be produced with the wave made up of signal elements consisting of half-cycle sine waves.

Squier System Applied to Submarine Cables. A more recent suggestion of Squier[5] gives the wave shown in Fig. 3*a*. This wave resembles the one advocated by Crehore and Squier in that each signal element consists of a half-cycle sine wave. As has been pointed out, there is no advantage gained by this.

The difference between the two systems lies in the fact that the wave in Fig. 3*a* uses three absolute values and crosses the axis once every half cycle. The code is the same as the Continental, a space being indicated by a half-cycle sine wave of one unit amplitude, a dot by a half-cycle sine wave of two units amplitude and a dash by a half-cycle sine wave of three units amplitude.

By referring to the figure, it will be seen that the resulting wave resembles a continuous sine wave, except for the fact that successive half cycles differ in magnitude. For this reason, the code may be termed an "unbroken-reversals" code.

In considering the application of this code to submarine cable telegraphy, it is convenient to make use of an analysis which is carried

[4] Crehore and Squier, loc. cit.

[5] Squier, loc. cit. *Proc. Phys. Soc.*

out in Fig. 3. Fig. 3*a* shows the words "United States" written in the code advocated by Squier. Fig. 3*b* shows a constant sine wave whose amplitude is equal to the amplitude of a dot in Fig. 3*a*. Fig. 3*c* shows the result obtained by subtracting the wave of Fig. 3*b* from the wave of Fig. 3*a*. On comparing this last wave with the wave

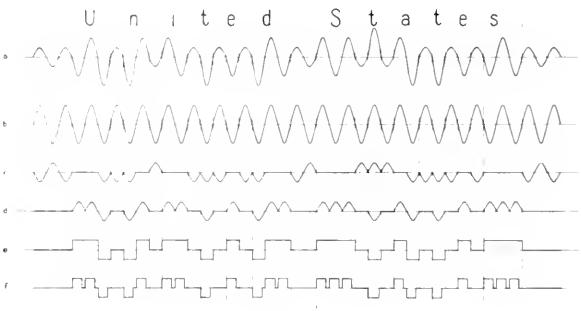


Fig. 3

a—Unbroken reversals code (space = 1 unit, dot = 2 units, dash = 3 units)
b—Constant sine wave, 2 units
c—Wave resulting when substracting *b* from *a*
d—Sine Wave code: note similarity between *c* and *d*
e—Rectangular wave, unmodified
f—Rectangular wave, modified by grounding apex one fourth of the marking time in addition to the spacing time

shown in Fig. 3*d*, it will be seen that the two waves are electrically equivalent. They differ only in having the signal elements permuted.

It is thus evident that the wave shown in Fig. 3*a* is made up of two components; one being the inert component shown in Fig. 3*b* which transmits no intelligence, and the other the intelligence carrying component illustrated in Fig. 3*c*.

The fact that the component shown in Fig. 3*b* does not carry intelligence from the sending station to the receiving station is made clear when we consider that its value at any moment is predictable and that the component can in fact be produced locally.

The net effect of this component is to reduce the voltage available for intelligence transmission to one-third of the total voltage. For example, if it is permissible to apply 60 volts to a particular cable, 40 volts out of these would be used up in transmitting the inert alternating-current wave and only the remaining 20 volts would be useful for the transmission of intelligence.

Radio and Carrier Telegraphy. Squier has also advocated[6] that the combination of sine wave envelopes, unbroken reversals and a three-current-value code be applied to radio and carrier telegraphy.

The advantages and limitations in applying codes with more than two current values have been fully discussed above, and do not need to be gone into further here. It will be evident that the combining with these of sine wave envelopes and unbroken reversals does no good.

The matter of using sine wave envelopes was discussed above, the discussion pointing out that waves with sine-wave envelopes are inferior to waves produced by sending rectangular shaped signals through suitable networks, both from the standpoint of the received signals, and from the standpoint of interference into other circuits.

The "unbroken reversals" bring in again the use of an inert component. Due to the fundamental difference between cable telegraphy on the one hand, and radio and carrier as usually practised on the other, the inert component in the latter case is somewhat smaller than in the former. In the code advocated by Squier, the current which may be subtracted without greatly affecting the intelligence-carrying capacity of the signals, is about one unit in value, which is the current corresponding to a space. When this current has been subtracted, the space current is reduced from one unit to zero, the dot current from two units to one, and the dash current from three units to two. This subtraction having been carried out, it is seen that the maximum intelligence-carrying component is approximately two-thirds of the maximum current actually employed. (This figure of two-thirds compares with the figure of one-third for the submarine cable.)

In the case of radio, the amount of power which must be radiated from the transmitting station is of particular importance. Since with the system advocated by Squier about two-thirds of the maximum voltage which is radiated is effective in transmitting intelligence, it is evident that about twice as much power must be radiated as would be required if the inert component were not transmitted.

Incorrect Assumptions. Two incorrect assumptions are made in the papers referred to and underlie a considerable portion of the arguments advanced in favor of the systems advocated by Squier.

One of these is that a wave, whose elements are half-cycle sine waves, lends itself to tuning. It is true that in the case of the "unbroken-reversals" code a certain amount of tuning can be secured, but this tuning applies only to the inert unvarying component in the wave, which carries no intelligence. The fact, shown in Fig. 2, that

[6] Squier, loc. cit., *Franklin Inst., Jl.*

the intelligence-carrying component contains no outstanding narrow range of frequencies to which tuning can be applied should make obvious the error in this assumption.

The other assumption is that a wave, which is ideal for the transmission of power, is also ideal for the transmission of intelligence. As a matter of fact, the transmission of intelligence inherently involves rapid and unpredictable changes in the current, whereas the transmission of power is best brought about by steady current, either direct or alternating. These two conditions are, of course, incompatible.

APPENDIX A

Use has been made of the following two principles:

1. In a telegraph circuit in which the line speed is near the maximum, the shape of the received dot is substantially independent of the shape of the impressed dot, and

2. The magnitude of the received current is approximately proportional to the area under the transmitted voltage curve.

The following general discussion of these principles has been furnished by J. R. Carson.

Let the arrival curve, due to suddenly impressed unit battery be denoted by $A(t)$; then the received signal $S(t)$, due to the elementary dot impressed signal $f(t)$ is given by[7]

$$S(t) = \int_0 f(x) A'(t-x)dx \tag{1}$$

the upper limit of integration being t for $t<T$ and T for $t \geqq T$. The latter case will alone be considered since the conclusions arrived at in this case are conservative.

Expanding $A'(t-x)$ in (1), we get

$$S(t) = \left[A'(t) - \frac{h_2 T}{2!} A''(t) + \frac{h_3 T^2}{3!} A'''(t) \ldots \right] \int_0^T f(x)dx \tag{2}$$

where $$h_2 = \frac{\int_0^T x f(x)dx}{T \int_0^T f(x)dx},$$

$$h_3 = \frac{\int_0^T \frac{x^2}{2!} f(x)dx}{T^2 \int_0^T f(x)dx}, \text{ etc.}$$

[7] J. R. Carson. "Theory of the Transient Oscillations of Electrical Networks and Transmission Systems." A. I. E. E. Trans., Vol. XXXVIII, 1919, p. 345.

It follows at once that, provided

$$\int_0^T f(x)dx \neq 0$$

and provided the duration T of the signal is sufficiently short, the arrival dot is given approximately by the leading term

$$A'(t)\int_0^T f(x)dx$$

and that this approximation becomes increasingly close as the speed of signaling is increased, *i.e.*, as the duration T of the dot is decreased.

The conclusions from the foregoing may be stated in the following propositions:

I. If the speed of signaling is sufficiently high the arrival signal representing the elementary dot is independent in shape of the form of the impressed signal, and is proportional in amplitude to the time integral or "area" of the impressed signal.

It will be evident, however, that if no restrictions are imposed on $A'(t)$ and $f(t)$, the foregoing proposition requires, in general, that the duration T of the dot shall be so small as to make the series expansion rapidly convergent from the start. This, however, requires a speed of signaling very considerably greater than that actually necessary in practise in order that the foregoing proposition shall hold to a good degree of approximation, at least for the types of impressed dot signals specially considered in the present paper. To show this, it is necessary to establish two less general propositions, valid for the types of impressed signals under consideration.

II. If the impressed signal $f(t)$ is everywhere of the same sign, then a value τ exists, such that $0 < \tau < T/2$, and such that

$$S(t+T/2) = A'(t+\tau)\int_0^T f(x)dx \qquad (3)$$

This proposition follows from the mean value theorem.

III. If $f(t)$ is everywhere of the same sign, and if further it satisfies the conditions of symmetry,

$$f(x) = f(T-x),\ (x \leq T/2)$$

then a value τ exists, such that $0 < \tau < T/2$ and such that

$$S(t+T/2) = 1/2\left[A'(t+\tau)+A'(t-\tau)\right]\int_0^T f(x)dx \qquad (4)$$

This last equation also follows from the mean value theorem. Furthermore, the conditions stated in proposition III are satisfied by

the rectangular wave, the half-cycle sine wave, and the rectangular wave extending through part of the dot provided the reference time $t=0$ is properly chosen.

Returning to proposition II, let us write

$$S_j(t+T\ 2)=A'(t+\tau_0+\tau_j)\int_0^T f_j(x)dx,$$

the subscript $_j$ indicating the particular type of impressed dot signal, and τ_0 the value of τ for any type of signal, taken as reference. Then

$$S_j(t+T\ 2)=\left[A'(t+\tau_0)+\frac{\tau_j}{1!}A''(t+\tau_0)+\ \ldots\right]\int_0^T f_j(x)dx \qquad (2a)$$

Now, the condition that proposition I shall hold to a good degree of approximation is that the expansion (2a) shall converge rapidly. Since the maximum possible value of τ_j is $T\ 2$ and since in practise it is much smaller than $T\ 2$, the required convergence obtains for much larger values of T, that is, slower speeds of signaling than that required in the expansion (1). Furthermore, for the three types of signals specifically under consideration τ_1, τ_2 and τ_3 differ from one another by quantities very much smaller than $T\ 2$ in all actual transmission systems.

If the conditions of proposition III are introduced, the approximation is still closer and proposition I is valid for still lower signaling speeds.

In order to arrive at quantitative ideas of the minimum signaling speeds at which the foregoing proposition is valid, it is necessary, of course, to specify the arrival curve of the transmission system under consideration. An application of the foregoing analysis to representative transmission systems both with and without a "cut-off" frequency has shown that it is valid to a very good degree of approximation for speeds considerably lower than the highest attainable under practical conditions.

APPENDIX B

Use has been made of the formula

$$H'=K\log m$$

where H' = the speed of transmission of intelligence

K = a constant

and m = the number of current values employed.

The assumptions which underlie this formula and its derivation will now be given.

Let us assume a code whose characters are all of the same duration. This is usually the case in printer codes. If n is the number of signal

elements per character, then the total number of characters which can be construed equals m^n. In order that two such systems should be equivalent, the total number of characters that can be distinguished should be the same. In other words,

$$m^n = \text{const.} \tag{1}$$

This equation may also be written

$$n \log m = \text{const.} \tag{2}$$

The speed with which intelligence can be transmitted over a circuit is directly proportional to the line speed and inversely proportional to the number of signal elements per character provided that the relations above are satisfied. Hence, we may write

$$W = s \cdot n \tag{3}$$

where s is the line speed. Substituting the value of n derived from the equation above, this equation becomes

$$W = \frac{s \log m}{\text{const.}} \tag{4}$$

which may also be written

$$W = K \log m \tag{5}$$

In applying this formula to practical cases it will be found impossible to comply strictly with the condition expressed by equation (1). As an example, consider the comparison between a three-current-value code where each character is made up of three signal elements, and a two-current-value code where each element is made up of five signal elements. It is obvious that the speed with which *characters* can be transmitted in the former case is five-thirds the speed in the latter case for a given line speed. In other words the ratio is 1.67:1 whereas the formula gives the ratio 1.58:1. It should be noted, however, that the former code possesses only 27 characters whereas the latter possesses 32. In other words one *character* of the latter code represents the transmission of more *intelligence* than one *character* of the former. Thus the figure 1.67 for the relative speeds of transmission of *characters* and the figure 1.58 for the relative speeds of transmission of *intelligence* are not incompatible.

It will be noted that the formula has been deduced for codes having characters of uniform duration and that it should not be expected to be anything but an approximation for codes whose characters are of non-uniform duration. To establish the formula for the latter case it would be necessary to make an assumption as to the relative frequencies of the various characters. It seems reasonable to sup-

pose that the formula will give a fair approximation to the facts in this case also, but it should not be expected to be accurate.

APPENDIX C

The deduction of the curves given in Fig. 2 from the curves given in Fig. 1 requires some explanation. Looked at casually, it would seem as if an isolated dot would not possess any frequency characteristics whatsoever. Nevertheless, if a voltage, such as any of those represented in Fig. 1, is applied to a network capable of being thrown into oscillation, the network will respond to the voltage by oscillating. Suppose, for simplicity, that the network consists of an inductance, a capacity and a very small resistance in series, the response of the network to the application of any of the voltages illustrated is that it oscillates at constant frequency and gradually decreasing amplitude. Further, the response varies when the natural period of the circuit is varied.

There are two ways of looking at this phenomenon. We may say, on the one hand, that the oscillations of the frequency in question are manufactured by the network out of the voltage applied and that the frequency does not exist in the original voltage. On the other hand, we may say that the original voltage contains components at or near the resonant frequency and that the circuit responds to these components, because it offers them a small impedance, while it does not respond to other components because it offers them a large impedance. Either of these views is permissible, but it is convenient for the purposes of this paper to use the nomenclature of the second view and to consider the applied voltages to be made up of an indefinitely large number of frequencies. The problem of determining the response of oscillating networks is then solved by deducing the frequency characteristic or the response characteristic of the impressed voltage. This characteristic may be determined by means of the Fourier integral, whose computation is described in any standard textbook on the subject. The following is intended to outline the considerations, from a physical standpoint, which lead to establishing this integral.

To deduce the frequency characteristic of an isolated dot, it is simplest to start with a long series of dots which are uniformly spaced. If such a series of dots is considered to extend indefinitely, it is possible to analyze the resultant wave into a Fourier series by well known methods. Now, suppose that such a Fourier series has been obtained for a given spacing of the dots. The next step is to increase

the spacing between the dots. The result of this is to increase the number of Fourier components in a given frequency range and to decrease the magnitude of each. If this process of increasing the space between the dots is continued indefinitely, we approach the condition of an isolated dot. Moreover, as we approach this condition, the number of components in a given frequency range increases indefinitely and the magnitude of each decreases indefinitely. This limiting result is known as the Fourier integral for the wave in question.

APPENDIX D

A table has been given in the paper in which the relative efficiency of various codes in transmitting intelligence is listed. The derivation of that table will now be given.

The comparison will include the following codes based on two current values: American Morse, Continental Morse, and the so-called "ideal" two-current-value code. It will also include the following codes based on three current values: Continental Morse and an "ideal" three-current-value code.

The assumption is made that the text is made up of five-letter-words, no allowance being made for punctuation. The following table gives the length of the spaces assumed in terms of signal elements.

	Ordinary Spaces Within Letters	Special Spaces in "Spaced" Letters	Spaces Between Letters	Spaces Between Words
American Morse (two current values)	1	2	3	4
Continental Morse (two current values)	1	—	2	3
Continental Morse (three current values)	—	—	1	2

It is assumed that the dashes in the two-current-value codes are of three signal elements duration, except for the letter *l* in American Morse which is assumed to occupy five signal elements. It may be that in practice, the dashes are somewhat shorter than has been assumed but the resulting error is not great. In connection with the relative spacings between letters and words assumed for the Continental and American Morse codes, it is also questionable whether they accord strictly with practise. It may be that these spacings are on the average more nearly equal than the table indicates. However, this assumption affects only the relative speeds obtainable with the American Morse and the Continental Morse and does not materially affect the comparison between codes based on two current values on the one hand and codes based on three current values on the other.

The term "ideal" has been applied to two codes which will next be explained. These codes are constructed on the same principles

as the Continental and American Morse codes with an effort to make them as brief as possible without making the reading too difficult. It is thought that the two ideal codes chosen are comparable in the matter of ease of reading. In constructing the two-element code, two steps are involved. In the first place it is assumed that the markings and spacings of any integral number of signal elements' duration can be used so that in addition to the values for markings and spacings assumed above, there may be dashes of two, four, etc., units duration. With these assumptions the 26 shortest characters that can be constructed are next made up. It is found that one character is of 1 unit duration, 1 of 2 units, 2 of 3 units, 3 of 4 units, 5 of 5 units and 9 of 6 units duration. The remaining 5 characters are taken of 7 units duration each. The second step is to ascribe the 26 letters of the alphabet to these characters in such an order that the most frequent letters correspond to the shortest characters. It is most efficient to use the same spacing as was assumed above for the Continental two-current-value code, with the addition that spaces of longer duration than three units may be employed within a letter.

The matter of constructing the ideal three-current-value code is similar. First, the 26 shortest characters are constructed. Two characters can be constructed having a duration of 1 unit, four characters having a duration of 2 units and eight characters having a duration of 3 units. The remaining twelve characters are taken 4 units in duration. Next, the most frequent letters are assigned to these characters in the order of their duration. It is best in this case to use the same assumptions as to spacings between letters and words as was used above in connection with the three-current-value Continental code. The use of spaces within letters is not economical in this case.

A frequency table given by Hitt [8] was used to determine the relative frequency of the various letters. The average duration per letter was computed from this table and corrected for spaces between words and letters. The resultant average duration is as follows:

Code	Signal Elements per Letter
American Morse (two current values)	8.26
Continental Morse (two current values)	8.45
Ideal (two current values)	6.14
Continental Morse (three current values)	3.77
Ideal (three current values)	3.63

[8] Parker Hitt. "Manual for the Solution of Military Ciphers." Army Service Schools Press, Fort Leavenworth, Kansas. Second edition, p. 7.

Abstracts of Bell System Technical Papers Not Appearing in the Bell System Technical Journal

The Auditory Masking of One Pure Tone By Another and Its Probable Relation to the Dynamics of the Inner Ear.[1] R. L. WEGEL and C. E. LANE. The authors used an air damped telephone receiver supplied with variable currents of two frequencies and determined the amount of masking by tones of frequency 200 to 3500 for frequencies from 150 to 5000. Except when the frequencies are so close together as to produce beats the masking is greatest for tones nearly alike. When the masking tone is loud it masks tones of higher frequency better than those of frequency lower than itself. If the masking tone is introduced into the opposite ear the effect occurs only by virtue of conduction through the bones of the head.

It is shown that combinational tones result when two tones of sufficient intensity are introduced simultaneously, these combinational tones being due to a non-linear response of the ear.

A dynamical theory of the cochlea is given which ascribes pitch discrimination to a passing of vibrations along the basilar membrane and a shunting through narrow regions of the membrane at points depending on the frequency. This view of the action of the ear offers an explanation of the masking effects.

Distribution of Radio Waves from Broadcasting Stations over City Districts.[2] RALPH BOWN and G. D. GILLETT. This is a description and analysis of the results obtained in a radio transmission survey of the cities of New York and Washington, D. C., and contiguous territory. Measurements of the field strength of radio signals from stations WCAP at Washington and WEAF at New York were made at a large number of points. Based on these data, curves are drawn showing how different kinds of territory cause different attenuations and showing radio shadows caused by mountains and by large masses of steel buildings. In order to visualize the phenomena, the data have also been plotted on maps, contour lines of equal signal strength being drawn. These contour maps illustrate graphically the non-uniformity of transmission in city areas and show the nature and extent of the "dead spots" and shadows.

[1] *Physical Review*, II, Vol. XXIII, p. 269, 1924.

[2] Presented to the Institute of Radio Engineers, February 16, 1924, at New York.

Measuring Methods for Maintaining the Transmission Efficiency of Telephone Circuits.[1] F. H. Best. The circuits involved in the transmission of speech in a modern telephone plant, particularly those designed for long distance operation, necessarily involve a considerable amount of complexity. The use of telephone repeaters the development of long toll cables, the application of carrier systems and other developments associated with these, while increasing the efficiency and economy of telephone toll circuits have also increased their complexity and have required the development of more effective means of insuring that the circuits are maintained at all times in good condition and adjustment.

Maintenance of the transmission efficiency of the telephone plant is conducted by a special force, using methods and apparatus that have been developed for this purpose. This paper gives a brief description of the transmission characteristics of some of the common type of telephone circuits, outlines a general method for measuring their transmission efficiency and describes several of the most modern types of transmission measuring sets, together with a brief, mention of the oscillators which supply the power for testing.

A Primary Standard of Light Following the Proposal of Waidner and Burgess.[2] Herbert E. Ives. The primary standard of light proposed in this paper consists of a black body constructed of platinum; the light from which, at its melting point, constitutes the photometric fixed point desired. The platinum black body consists of a cylinder of highly polished platinum with a narrow slit for observing the interior. Studies of the optical properties of reflecting cylindrical enclosures show that at certain angles of observation the interior is practically "black." The platinum cylinders are heated electrically and the light from the interior is observed by throwing an image of the slit on to a photometer field. Two series of observations were made, one by a visual photometric method, the other by a photoelectric cell giving a photographic record by means of a string electrometer. The two methods of observation gave practically identical results, yielding a final value for the brightness of the black body at the melting point of platinum of 55.4 candle power per square centimeter. The advantages of this proposed standard over the present unsatisfactory flame standards are discussed.

High Quality Transmission and Reproduction of Speech and Music,[3] W. H. Martin and Harvey Fletcher. Radio broadcasting has

[1] Journ. A. I. E. E. Vol. XLIII, p. 136, 1924.
[2] *Journal Franklin Institute*, Vol. 197, p. 147, p. 359, 1924.
[3] Journ. A. I. E. E. Vol. XLIII, p. 230, 1924.

drawn attention to the problems involved in obtaining high quality in systems for the electrical transmission and reproduction of sound. This paper gives the general requirements for such systems, discusses briefly the factors to be considered in design and operation and indicates to what extent the desired results can be obtained with the means now available.

It was pointed out in this paper that broadcasting stations and connecting lines can be made practically perfect but that most of the loud speaking apparatus now extensively used for reproduction, causes distortion. At the time of reading this paper the authors demonstrated a laboratory model of a new loud speaker of unusual design. This apparatus reproduces all frequencies from the lowest to the highest of the audible range with approximately equal facility. This results in reproduced music which the ear can scarcely distinguish from the original.

Telephone Transformers.[1] W. L. CASPER. After outlining the varied sets of conditions which different types of telephone transformers must meet, this paper discusses the design and construction of transformers to handle efficiently the range of frequencies ordinarily present in speech. Two winding transformers only are dealt with, and the three most common impedance combinations of the two circuits connected by the transformer are considered; namely, both circuits comprised of resistances, one circuit a resistance, and the other a positive reactance, and one circuit a resistance and the other a negative reactance.

The efficiency with which energy is transmitted is measured by comparison with an ideal transformer, and the transformer is studied by supposing it replaced by an equivalent T network. The variation of transformer losses with frequency is discussed and characteristic curves are shown for transformers of different mutual impedances. Characteristics are also given showing the operation of the in-put transformer associated with the vacuum tube.

The mechanical construction of the common battery repeating coil, telephone induction coil, and of certain types of transformers for vacuum tube circuits, are shown. These transformers are all constructed so as to give the desired accuracy of speech transmission under their respective circuit conditions.

Radio Telephone Signaling—Low Frequency System.[2] C. S. DEMAREST, M. L. ALMQUIST and L. M. CLEMENT. The system described

[1] Journal of the American Institute of Electrical Engineers, Vol. XLIII, p. 197, 1924.

[2] Journ. A. I. E. E. Vol. 43, p. 210, 1924.

provides a means whereby any one of about seventy-five radio stations, operating on the same wave length, may be called without signaling the remaining. Obviously this is an important improvement in the radio art for in many cases it permits a radio station operator to pursue other duties which would be impossible if he were required to listen in at all times.

The engineering problem presented, being remarkably similar to many telephone problems, was solved in a very similar manner. When it is desired to signal a station, an alternating current of a very definite frequency is impressed on the transmitter. This modulates the power radiated similar to the way the undulations of the voice modulate the power when speech is transmitted. The station to be signaled is determined by the code transmitted. This code consists of a definite grouping of dots and spaces and dashes.

At the receiving station this modulated power is detected in the usual manner and results in an alternating current identical in nature to that used in transmitting the code. A special alternating current relay of high selectivity and sensitivity, in conjuction with a more common direct current relay system, converts the code into a series of direct current impulses. These impulses pass into a selector like that used in common train dispatching circuits. The mechanism of this selector will be unlocked and a local ringing circuit closed if the code is that for which it has been set. Thus it is seen that the code is received by all stations but only one selector of the system will operate to ring its local annunciator bell. The number of stations which can operate in the same system is determined by the number of possible combinations on the selector. At present this is set at seventy-eight but this may be readily extended to include more than two hundred.

Because of the high selectivity of the alternating current relay and its associated direct current relay system, the apparatus is particularly free from interference such as the operation of nearby spark or I.C.W. Stations. In fact, tests show that the signaling system will continue to function satisfactorily long after interference is so bad as to make conversation impossible. As designed, the signaling system may be made an integral part of a standard radio system without altering the apparatus already in use.

Contributors to this Issue

H. R. Friis, E.E., Royal Technical College in Copenhagen, 1916. Columbia University, 1919–1920. Research Department, Western Electric Company, 1920—. Mr. Friis' work has been largely in connection with radio reception methods and measurements. He has published papers on vacuum tubes as generators, radio transmission measurements and static interference.

A. G. Jensen, E.E., Royal Technical College of Copenhagen, 1920. Research Assistant to Professor P. O. Pedersen, 1920–21. Columbia University, 1921–22. Research Department, Western Electric Company, 1922—. Mr. Jensen has been mainly engaged in work relating to radio reception methods and measurements.

D. D. Miller, B.S. in electrical engineering, Tennessee, 1909. Installation Department, Western Electric Company, Hawthorne, 1909–1910. Physics Laboratory, Engineering Department, New York, 1910–1917. Apparatus Development, 1917—. Mr. Miller is in charge of the design of relays and has contributed much to the development of the modern flat types of relays which combine cheapness of manufacture with improved operating characteristics.

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.

C. F. Sacia, B.E.E., University of Michigan, 1916; Engineering Department of the Western Electric Company, 1916—. Mr. Sacia has been engaged upon methods for recording and analysis of speech.

E. B. Wheeler, B.S., University of Illinois, 1905. Engineering Department, Western Electric Company, Chicago, 1905–1907. Engineering Department, Western Electric Company, New York. Physical Laboratory, 1907–1921. General Development Laboratory, 1921—. Mr Wheeler has been actively connected with the development of improved types of switchboard and telephone cords, dry

cells, condensers, and other types of telephone equipment; and with the investigation of the effects of atmospheric conditions upon the performance of telephone apparatus.

Ronald M. Foster, S.B., Harvard, 1917; American Telephone and Telegraph Company, Engineering Department, 1917–19; Department of Development and Research, 1919—.

Karl K. Darrow, S.B., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., in physics and mathematics, University of Chicago, 1917; Engineering Department, Western Electric Company, 1917—. At the Western Electric, Mr Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

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.

Charles R. Moore, B.S. in Mechanical and Electrical Engineering Purdue, 1907; E.E., Purdue, 1910; Instructor and Assistant Professor Electrical Engineering, Purdue, 1907–13; Manager of LaFayette Electric and Mfg. Co., 1913–14; Associate in Electrical Engineering, University of Illinois, 1914–16; Engineering Department of the Western Electric Co., 1916—. At the Western Electric, Mr. Moore, for several years, has had charge of transmitter development work and has contributed important inventions relating to telephone instruments and acoustic devices.

Harry Nyquist, B.S. in electrical engineering, North Dakota, 1914; M.S., North Dakota, 1915; Ph.D., Yale, 1917; Engineering Department, American Telephone and Telegraph Company, 1917–1919; Department of Development and Research, 1919—. Mr. Nyquist has been engaged in work on both direct current and carrier telegraph systems as well as problems in line compositing.

The Bell System Technical Journal

July, 1924

Electrical Tests and Their Applications in the Maintenance of Telephone Transmission

By W. H. HARDEN

Introduction

THE installation and maintenance of the circuits in a telephone plant employed for the transmission of speech require the use of various testing schemes to insure a high grade of commercial service. Circuits are engineered and installed to meet the established standards of transmission in the most economical manner and this having been done the next step is to provide an adequate testing program. A number of the electrical tests required in this program include well known laboratory methods adapted so that they can be readily applied in the field, while others have been developed for particular use in telephone maintenance work.

Standard types of test boards and portable testing arrangements are as a rule made up of simple circuits designed electrically and mechanically in a manner to facilitate ready connection to the operating circuits in the plant. It has been found by experience that many of the transmission maintenance requirements can be taken care of by direct current testing methods and the simpler alternating current tests. With the advent of vacuum tubes, some of the more complex circuits such as repeaters and carrier called for the development of testing apparatus to meet the additional maintenance requirements. Fortunately, the vacuum tube furnished the means whereby new testing devices have been provided which can be applied as quickly and readily to maintenance work as the simpler methods.

In what follows is given a discussion of the more important electrical testing methods together with the application of these methods in maintaining the transmission efficiency of the various types of telephone circuits now in general use. Direct current testing methods are covered first and later alternating current methods are considered. A typical toll connection is used to illustrate the general scheme of applying the various electrical tests in everyday installation and maintenance work.

Direct Current Tests

The tests involving the use of direct currents and voltages provide means for checking some of the electrical characteristics of telephone circuits and insuring to a certain extent that these circuits will give satisfactory speech transmission. The application of these tests to the telephone plant reduces to a minimum the amount of alternating current testing required and lengthens the interval at which alternating current tests need be made.

Wheatstone Bridge Measurements. The various arrangements of the Wheatstone bridge for direct current measurements and the principles involved are well known and are therefore not discussed in any detail in this paper. However, due to the importance of such measurements in the maintenance of telephone circuits and in trouble location work a brief discussion of the general applications of the bridge is given.

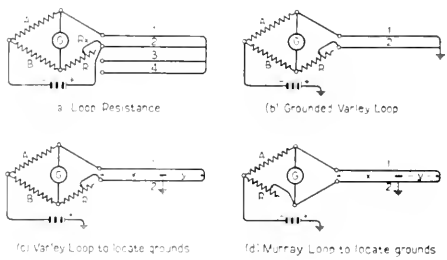


Fig. 1

Fig. 1 shows four arrangements commonly employed in routine testing and trouble location. Diagram (a) of this figure gives the bridge circuit for obtaining loop and single wire resistance measurements. Diagram (b) shows the circuit for Varley loop measurements to determine resistance unbalances in a pair when a third wire is not available, while Diagrams (c) and (d) show the Varley and Murray bridge circuits used in locating grounds. Various other arrangements of the bridge circuit are, of course, used where such arrangements will facilitate the testing work. For a condition of bridge balance indicated by no deflection of the galvanometer the value of the resistance being measured is given by the well known bridge ratio formulae. In diagram (a) of Fig. 1, for example, $R_x = \frac{A}{B} R$.

The testing circuits shown in Fig. 1 are commonly used in the day by day maintenance of the telephone plant. Resistance and resistance balance measurements are made periodically on toll circuits to guard against series resistance unbalances such as might be caused by high resistance joints. The Varley and Murray tests are constantly employed in directing linemen in clearing trouble such as crosses and grounds. The Wheatstone bridge is therefore an important feature of toll test boards where keys are provided to furnish a means for quickly setting up the different bridge test circuit arrangements desired.

The Varley or Murray tests used in connection with pole line diagrams in locating troubles provide a means whereby the test board men can direct the movements of linemen to the best advantage. Unit resistance values with temperature corrections are available for different types of circuits. If a good circuit of the same type and gauge over the same route is available, the unit resistance can be determined directly by a loop measurement of this circuit. The resistance values obtained by measurements on circuits having crosses or grounds can then be used to determine the distance to the trouble and the lineman sent to this point. By making measurements carefully and using the most accurate unit resistances available, troubles can be located and cleared in the minimum amount of time. In trouble location work on cables where the cable needs to be opened to repair the trouble, bridge measurements are made to give the approximate distance to the fault. More exact locations can then often be made by using an exploring coil test set by means of which the cable repairman listens by induction to a tone sent out from the cable terminal and determines in this way when he passes the point of trouble.

Leakage or Insulation Resistance Measurements. An important factor in the maintenance of telephone circuits is to insure that there are no resistance leaks between conductors or between conductors and ground. It is also important to insure that insulated conductors will not have the insulation broken down by the voltages which are met with under service conditions. Two types of tests now used extensively in the plant are described below:

(1) *Voltmeter Method.* This method is the one commonly used in determining the leakage between wires and to ground particularly on toll circuits involving open wire and on subscribers' circuits. As shown in Fig. 2 the testing arrangement consists of a voltmeter in series with a battery connected to the conductors under test. Diagram (a) shows the connection for testing the leakage between wires and

Diagram (b) the connection for ground leakage tests. Since the leakage resistance measured is relatively high, the most accurate results are obtained by using a high resistance voltmeter and a fairly high test voltage. In practice, a 100,000-ohm voltmeter is generally used with a test battery of from 100 to 150 volts. A test voltage of 200 is also provided in circuit with a milli-ammeter and protective resistance for use in checking the strength of insulation of central office wiring and subscriber's lines.

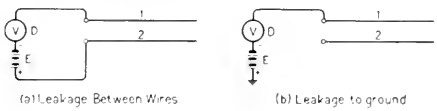


Fig. 2

Considering the circuits shown in Fig. 2, the voltage of the battery E is equal to the IR drop over the voltmeter plus the IR drop or the drop due to leakage over the remainder of the circuit back to the battery. Designating the insulation resistance being measured by X, the voltage of the test battery by E, the deflection in volts of the 100,000-ohm voltmeter by D and the current flowing by I, then

$$E = D + XI,$$

$$D = 100{,}000\ I,$$

and

$$X = 100{,}000 \left(\frac{E - D}{D} \right).$$

In practice, tables are provided from which the insulation resistance or leakage can be read directly for various deflections of the voltmeter. When expressed in terms of insulation resistance, the most convenient unit of measurement is the megohm. If expressed in terms of leakage, the unit used is a reciprocal function of the megohm known as the millimicromho. The results of measurements for complete circuits are generally reduced to apply to a unit length of circuit such as a mile so that the testing results on circuits of different lengths will be comparable.

The open wire toll circuits in the telephone plant are tested periodically by the method just described. The leakage of circuits is materially increased by defective or broken insulators and by contact of the wires with foreign objects such as trees, particularly under damp weather conditions. Troubles of this kind are detected by careful

leakage measurements and routine tests, therefore, become very useful in indicating when remedial measures, such as line inspections and tree trimming work, should be undertaken.

If open wire telephone circuits are so situated that contact with foliage growth will occur during the growing season low values of insulation resistance are certain to result even under dry weather conditions. This is illustrated by the curve of Fig. 3 which shows

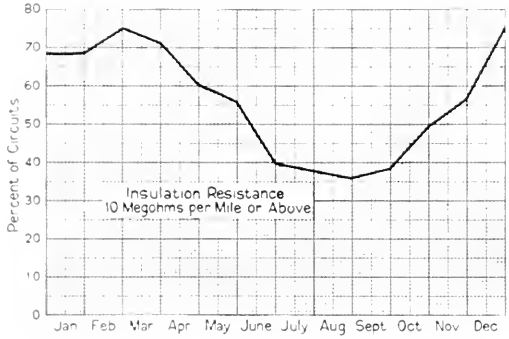


Fig. 3

results of monthly day time dry weather insulation measurements on a number of toll circuits over a period of a year under conditions of this kind. The monthly testing periods are plotted as the abscissa while the ordinates show the percentage of circuits which measure 10 megohms per mile or more during these monthly testing periods. This curve indicates the need for periodic insulation resistance tests and the use which can be made of such tests in instigating clean-up work.

2. *Megger Method.* The voltmeter method is not applicable for accurately testing the higher values of insulation resistance such as are encountered in telephone cables. Conductors in cables require a very high insulation and in practice values of 500 megohms or more per mile are specified. The laboratory galvanometer method of testing very high resistances, which is the same in principle as the voltmeter method, can, of course, be used, but is not sufficiently rugged for field testing. To take care of cable testing work in the plant a method known as the Megger method is employed. Fig. 4 gives the

schematic circuit arrangement of the "Evershed" megger which is the commercial form of instrument now used for this work. This circuit is contained in a small portable box so that it can be readily used at office frames or carried out on the line.

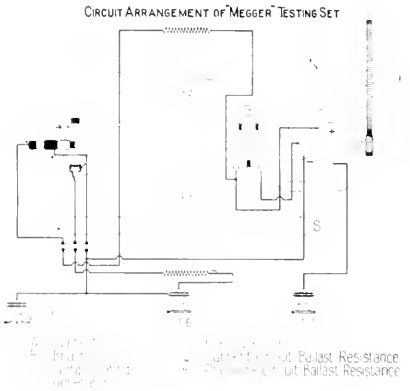


Fig. 4

The circuit of Fig. 4 consists of a high potential direct current generator, hand driven by means of a crank handle. This generator is arranged to provide a maximum potential of 400 volts which is the potential now employed in measuring insulation resistance on cable conductors. The potential furnished by the generator is impressed on two coils in an indicating device, one of these coils being in series with a fixed resistance and the other in series with the circuit under test. When the circuit under test is not connected to the megger the full current from the generator flows thru the first coil of the indicating device which for this condition causes the pointer to go to the "Infinite" position. When a circuit is connected to the megger "Line" and "Earth" terminals some current flows through this circuit due to its leakage and through the second coil of the indicating device. This causes the indicating device to move the pointer toward the "Zero" position, the amount of the movement depending on the leakage in the circuit under test. The indicating device is calibrated in megohms so that the results of insulation tests can be read directly.

Current Flow and Voltage Measurements. Tests to determine the amount of direct current flowing in telephone circuits involve the simple arrangement of an ammeter or milli-ammeter in series with a d.c. generator or battery and the circuit under test. The amount of current flowing is, of course, a function of the resistance of the circuit and the voltage applied. In testing arrangements where it is neces-

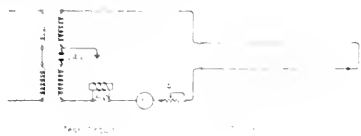


Fig. 5

sary to know the battery or generator potential, voltage readings are made by using ordinary voltmeters having the proper range and resistance. Direct current and voltage measurements can best be described by considering two of their applications in the telephone plant.

Fig. 5 shows a simple test circuit provided in the local test desk whereby central office battery is supplied through a regulating rheostat, a standard cord circuit and a meter. Knowing the voltage of the central office battery and the resistance in the test circuit, the reading of the meter when circuits such as a subscriber's loop or trunk conductors are connected and shorted at the distant end gives a means for determining the direct current resistance of these. Tables are generally provided for use at the test desks by means of which different readings of the meter for different conditions of measurement can be converted directly into resistance values. The rheostat in the test circuit is provided primarily for adjusting the current supplied to subscribers' loops and instruments to the same value for different lengths of loop. Talking tests as mentioned later in connection with substation maintenance can then be made from the instruments to the test man in the central office under the same current supply conditions for different lengths of loop at the time substations are installed or when these are reported in trouble. The arrangement shown in Fig. 5 is useful in detecting high resistances in circuits when a Wheatstone bridge is not available. High resistances in the main frame protector springs and heat coils of both subscribers' lines and toll circuits are also determined by a current flow method, a special portable testing set, however, being designed particularly for this purpose.

Direct currents and voltages are very important factors in the operation and maintenance of amplifier circuits such as telephone repeater and carrier apparatus. The battery supply arrangements for a single tube amplifier are shown in Fig. 6.

It is necessary in order to insure efficient amplification without distortion to regulate the currents and voltages to fairly close limits. In practice provision is made for quickly reading the voltages of grid, filament and plate batteries as shown by the voltmeter connection (V) in the figure. The plate current is read by the milli-ammeter M and the filament current by the ammeter A. The filament current

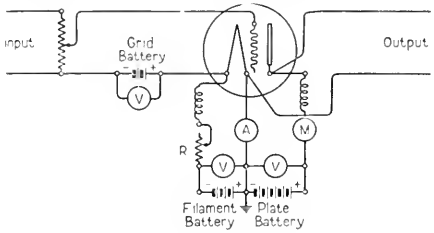


Fig. 6

is regulated to meet the operating limits by cutting resistance in or out of the circuit with the rheostat R. The same applications of current and voltage readings apply to the more complicated amplifier circuits, although wherever practicable automatic regulating devices are provided which reduce the amount of manual testing work to a minimum.

Capacity Measurements. There is little occasion in transmission maintenance work to make accurate direct current measurements of capacity. A simple d.c. test, however, has been provided for use primarily on subscribers' loops for checking the condensers in the sets.

Fig. 7

As shown in Fig. 7 the circuit consists of a 100,000-ohm voltmeter in series with a grounded 100-volt battery connected to one conductor of a subscribers' loop, the other conductor of the loop being grounded.

When the battery is connected a current will flow momentarily in the circuit charging the condenser C. This will produce a throw of the voltmeter needle, the amount of the deflection depending upon the capacity of the condenser C and the capacity between conductors. If the tip and ring connections of the loop are reversed the volt-

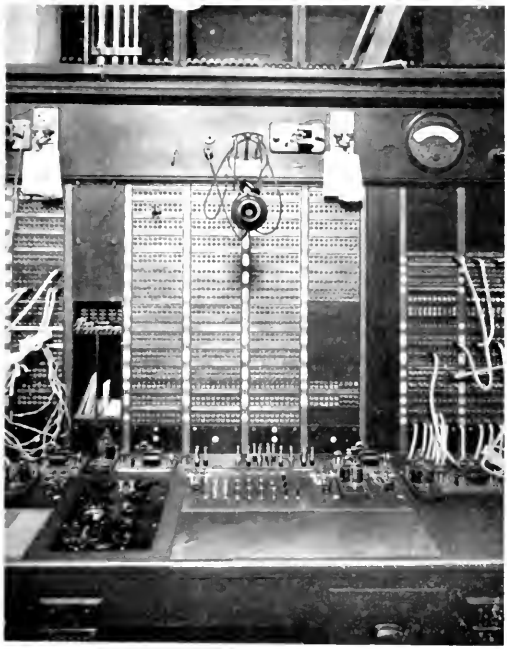


Fig. 8

meter needle throw will be in the opposite direction from that obtained in the first condition. The capacity of the conductors in the loop is relatively small as compared to that of the condenser C so that by knowing the throw which should be obtained under the test conditions for known values of capacity a fairly good check of the condensers in subscribers' sets is provided by this method of measure-

ment. Different deflections of the voltmeter needle will, of course, be obtained depending on whether the loop tested is a single party, two-party or four-party line and also on whether 1 μ f. or 2 μ f. condensers are provided in the substation sets. These conditions must be known by the testman if he is to properly interpret the testing results and detect missing or defective condensers.

Standard Types of Testboards. Pictures of two of the latest types of toll and local testboards are shown in Figs. 8 and 9. These boards provide circuit arrangements for making most of the direct current

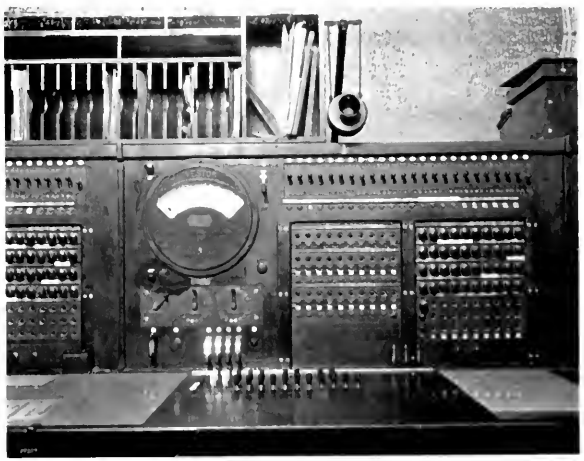


Fig. 9

tests which have just been described and also some of the alternating current tests described below. The wiring of the test circuits to keys, jacks and plugs and the provisions made for picking up various telephone circuits for test greatly facilitate routine maintenance work and the location of troubles which occur in service. Modifications and various arrangements of the tests described above have been provided for in these boards to meet different operating conditions which may arise.

The test board shown in Fig. 8 is designed primarily for testing toll circuits. The vertical section of the board provides jacks for

terminating the toll circuits and the apparatus associated with them, such as phantom and simplex coils, composite sets, etc. The 100,000 ohm voltmeter and the Wheatstone bridge and keys for obtaining various testing arrangements are mounted in the horizontal shelf and connections are made to the toll circuit and equipment jacks by means of the cords and plugs located at the back of the shelf. The telegraph instruments are used on order wires to distant test boards and the meter shown in the vertical section of the board is for measuring the voltage and current in telegraph circuits.

The test board shown in Fig. 9 is designed primarily for testing the local plant, although tests on toll circuits can also be made from this board. One transmission feature provided in the board is an artificial line which when cut in circuit with a 500 ohm subscriber's loop, gives an overall equivalent of approximately 30 TU. This line is terminated on keys by means of which it can be connected as a trunk circuit and used in talking tests on subscribers' loops at the time of their installation or when subscribers' stations are visited in connection with trouble complaints. Jacks are provided in the vertical section of the board for terminating certain test trunks and other test trunks are terminated on keys. A Wheatstone bridge is not normally mounted in this type of test board, but where required, a portable bridge is supplied which is generally kept in one of the drawers of the board when not in use.

Alternating Current Tests

While the direct current tests just described tell a great deal about the physical and electrical condition of telephone circuits, it is very necessary in maintenance work to consider also the alternating current characteristics. The transmission of speech is, of course, fundamentally a problem of the transmission of alternating currents of very small values. The inductance and capacity as well as the resistance and leakage of circuits, therefore, become important items in determining the efficiency of telephone circuits and means must be provided for testing these characteristics under operating conditions. In principle, alternating current testing methods do not differ materially from direct current methods and their application in the telephone plant is not difficult.

Alternating Current Bridge Measurements. These measurements employ Wheatstone bridge arrangements, the direct current source of power being replaced by an alternating current source and the condition of bridge balance being obtained by some alternating

current detecting device, generally an ordinary telephone receiver. Four important bridge measuring methods are used extensively in telephone testing work as described below:

(1) *Alternating Current Capacity Tests.* The bridge circuit arrangement for measuring a.c. capacity is shown in Fig. 10.

Two arms of the bridge consist of fixed and equal resistances A and B connected by a slide wire resistance, the position of the contactor on this slide wire determining the total amount of resistance

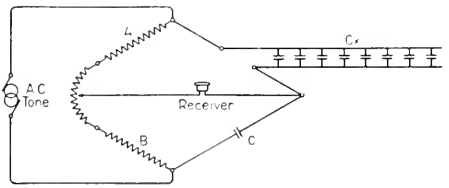


Fig. 10

in each of the two arms. The fixed resistances in A and B are simply extensions of the slide wire and can be cut out of the circuit when not required. The third arm of the bridge consists of standard condensers C, and the fourth arm the circuit whose capacity C_x is to be measured. A source of alternating current generally an 800 or 1,000 cycle oscillator is connected to the terminals of the arms A and B while the telephone receiver is connected to the slide wire contactor and to the junction of the standard condenser and circuit under test. A balance of the bridge is obtained when there is minimum tone in the receiver, for which condition the common bridge formula $C_x = \frac{B}{A} C$ applies. The slide wire is calibrated to read the ratio $B \div A$ directly.

For field testing work the above circuit arrangement is made up in a portable box and a portable oscillator is used so that the apparatus can be readily carried about as required. The commercial form of bridge provides three values of standard condensers which can be used to cover measurements from about 500 micro-microfarads up to 1.5 microfarads. This bridge finds its application in the plant in measuring the capacity of short lengths of non-loaded cable, bridle wire, switchboard wire, etc. Such measurements are of particular importance in connection with the installation of 22 type telephone repeaters to determine the proper values of building out condensers to use in the line and balancing circuits.

Another use which is made of alternating current capacity measurements is in connection with the open location test provided at toll test boards. The essential features of the circuit arrangement are shown in Fig. 11.

The ordinary Murray connection of the test board bridge is used, the four arms of the bridge consisting of one fixed 1,000 ohm resistance A, a variable resistance R, a standard 1 μf. condenser C and the open

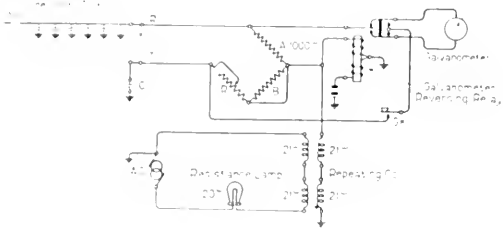


Fig. 11

condenser under test C_x. Ordinary 20 cycle ringing current is used as the measuring current and the galvanometer or voltmeter connected through a reversing relay so that it will always read in one direction. For the balanced condition of the bridge as indicated on the galvanometer the relation $C_x = \frac{R}{A} C$ holds. Substituting the numerical values for A and C in the above formula C_x then equals $\frac{R}{1000}$.

The above test provides a means for determining the approximate distributed capacity of a circuit up to the point where it is open. With previous measurements on known lengths and similar types of circuits available and assuming the distributed capacity proportional to the length of circuit, this test provides a simple means for determining the approximate distance out to the open. In practice fairly good results are obtained on loaded or non-loaded open wire circuits up to 200 miles in length and on loaded or non-loaded cable up to 40 miles in length. The degree of accuracy with which opens can be located by this method depends, of course, on having good unit capacity measurements for the different types of circuits involved in the testing work.

(2) *Capacity Unbalance Tests.* If the electrostatic capacities between wires and between wires and ground in telephone circuits are not

properly balanced crosstalk between circuits will result. The effects of capacity unbalances of this kind are particularly serious in producing side to side and phantom to side crosstalk in quadded cable circuits unless great care is taken in splicing the various pairs and quads in consecutive lengths so that the resultant unbalances will be a minimum. This is to be expected since in cables the electrostatic capacities between conductors and between conductors and sheath

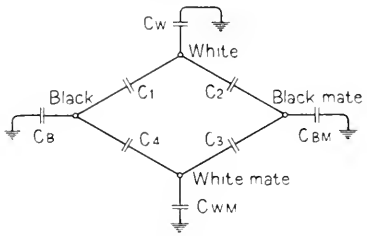


Fig. 12

are high as compared to open wire circuits and any irregularities in construction may produce very appreciable unbalance conditions between these capacities.

Fig. 12 shows the direct electrostatic capacities in a quad which, if they do not have the proper balance relations will produce excessive crosstalk. The conductors of one pair are designated "white" and "white mate" and of the other pair "black" and "black mate." The particular arrangement of the conductors in the figure to form the arms of a Wheatstone bridge is used since this arrangement is employed in the capacity unbalance measuring circuit described later.

Neglecting second order effects, side to side crosstalk is produced by unbalances in the direct capacities between conductors in accordance with the following relation.

$$\text{Capacity unbalance} = C_1 + C_3 - (C_2 + C_4).$$

In phantom to side crosstalk the unbalance relations of the direct capacities of the conductors to ground (sheath and "bunch") in addition to the direct capacities between conductors become important. Again neglecting second order effects, the unbalance relations producing crosstalk between the phantom and the "white" side is

$$\text{Capacity unbalance} = 2\left[(C_1 + C_2 - (C_3 + C_4) + \tfrac{1}{2}(C_H - C_{WM})\right].$$

Similarly the unbalance relations producing crosstalk between the phantom and the "black" side is

$$\text{Capacity unbalance} = 2\left[(C_1 + C_4) - (C_2 + C_3) + \tfrac{1}{2}(C_B - C_{BM})\right].$$

The factor 2 enters into the last two formulæ since the difference in direct capacities have about twice the effect on phantom to side crosstalk as they do on side to side crosstalk.

The capacity unbalances given above are measured on each quad in every loading section and give a measure of the side to side and phantom to side crosstalk due to capacity unbalance in the cable. Such measurements are usually made at three points in every loading section and the quads are spliced at these points in such a way that the capacity unbalances in the two directions will tend to neutralize. In this connection particular care is taken to neutralize the phantom to side unbalances since these are usually higher.

For making capacity unbalance tests a special portable bridge known as the capacity unbalance test set was developed which has been in general use since the introduction of quadded cables in the telephone plant. Fig. 13 shows the schematic circuit arrangement of this bridge for measuring the capacity unbalance as indicated above between sides of a quad.

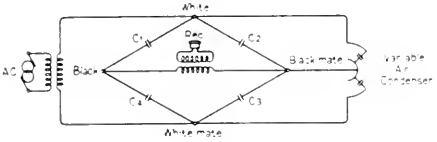


Fig. 13

The two conductors of each side circuit of the quad are connected to opposite corners of the bridge, these being designated as "white and "white mate" and "black" and "black mate." The direct capacities between these conductors then become the arms of the bridge. An oscillator is connected through a transformer to the "white" and "white mate" terminals of the bridge and a variable air condenser is connected to these same terminals. A telephone receiver is connected through a transformer to the "black" and "black mate" terminals. The variable air condenser is adjusted until a minimum tone is observed in the receiver, this adjustment adding capacity to one side or the other of the bridge. The variable condenser is calibrated to read the unbalances directly in micro micro-

farads, the direction of the unbalances being indicated by red and black scales and arbitrarily designated as (+) and (−).

Fig. 14 shows the circuit arrangement of the bridge for measuring the capacity unbalance between the phantom and "white" pair. The oscillator, variable condenser and receiver are connected as before, the "black" conductor and its mate however, being strapped together at one of the remaining bridge terminals and ratio arms R_1 and R_2 each consisting of 2,000 ohms resistance, being connected as shown to the fourth bridge terminal. For the condition of minimum tone the variable condenser reading then gives a measure of the capacity unbalance between the phantom and the "white" pair, that

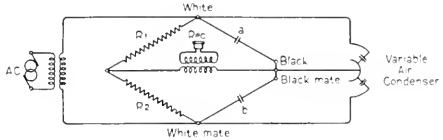


Fig. 14

is $a-b$. The capacities a and b take into account in this case the capacities of the "white" and "white mate" conductors to ground in addition to the direct capacities between wires shown in Fig. 13. The unbalance between the phantom and "black" pair is obtained in the same manner as shown by interchanging the "white" and "black" conductor connections to the bridge. The test set reads only half the capacity unbalance as defined in the above formulæ for phantom to side unbalance.

In practice the testing arrangement just described is used to test unbalances of all quads in a cable in each direction. At any splicing point where the tests are made the three unbalance measurements in each direction for each quad are carefully recorded and the splices then made by combining (+) and (−) values so as to neutralize each other as much as possible thereby reducing the resulting capacity unbalances and the crosstalk in each direction to a minimum. Both the bridge and oscillator are readily portable and designed for outdoor use. The bridge is equipped with keys, binding posts and leads to allow connections to be quickly made to the cable conductors and the various conditions of unbalance measured.

(3) *Impedance Tests.* The various bridge arrangements for capacity measurements are essentially impedance measuring devices, the im-

pedance of condensers being negative reactance. In telephone circuits and equipment where inductance is involved such as in loading coils, transformers, retardation coils, etc., the effective resistance as well as the inductance becomes a factor which must be taken account of in bridge testing work. For measuring effective resistance, inductance and impedance, bridges have been developed which are similar to capacity bridges except that standard condensers in the balancing arm are replaced by standard inductances and resistances.

There are two general types of bridges in use in the telephone plant designed to measure impedance; one type for testing equipment made up mostly of inductance, such as loading coils, and the other for testing the impedance characteristics of various types of equipment and circuits generally within the operating range of frequencies.

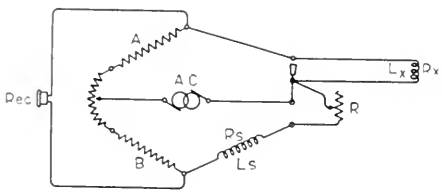


Fig. 15

The circuit arrangement shown in Fig. 15 is for an impedance bridge designed primarily for measuring impedance of equipment having positive reactance characteristics. As in the capacity bridges two arms are made up of fixed resistances *A* and *B*, connected by a slide wire resistance. The impedance to be measured makes up the third arm of the bridge and the standard impedance consisting of known values of inductance and resistance is the fourth arm. To obtain accurate measurements requires that the standard impedance be approximately the same order of magnitude as the impedance measured and the phase angles of the two must be very nearly the same. Values of standard inductance are, therefore, chosen which are known to be fairly near the values of the unknown inductances and a variable resistance *R* is provided which can be switched in series with either arm of the bridge and adjusted until the resistance components in the two arms are equal. The bridge is balanced by adjusting the slide wire resistance and the resistance *R* until a mini-

mum tone is heard in the receiver. For the condition shown in Fig. 15 when the bridge is balanced $L_x = \frac{A}{B} L_s$ and $R_x = \frac{A}{B}(R+R_s)$. The slide wire is calibrated to read the ratio $\frac{A}{B}$ directly and tables of values for L_s and R_s at various frequencies are supplied for use with the commercial form of bridges. The value of R is read directly from the dial rheostats on the bridge.

In practice this form of bridge finds its principal application in measuring the inductance and resistance of cable loading coils when trouble is experienced which necessitates opening up the cable and loading coil pots. It is also used to measure the unbalance between windings of coils as, for example, between the line windings or the drop windings of repeating coils. For measurements of the latter kind one winding is connected in place of L_s and the other in place of L_x and the unbalance between the two windings is then given by the slide wire ratio. A further use of this scheme is in checking the correctness of loading of short cable circuits and a special bridge has been designed for this purpose. A pair which is known to be properly loaded is used as the standard and all other pairs of the same length and loading are checked by connecting them one at a time into the unknown arm of the bridge.

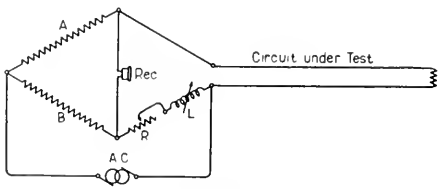


Fig. 16

The form of bridge designed to measure the impedance characteristics of circuits and equipment at any desired frequency or at a number of frequencies is shown in Fig. 16.

The fixed resistances A and B, generally of 1,000 ohms each, make up two arms of the bridge, the circuit under test the third arm and a variable resistance and a variable inductance standard the fourth arm. The variable inductance L is arranged so that it can be switched in series with the circuit under test when the characteristics of this circuit are such that its capacitive reactance predominates. For a

condition of balance indicated by minimum tone in the receiver, the effective resistance of the circuit is given directly by the value of the variable resistance R, and the inductance by the value of L. For any particular frequency f at which a measurement is made, the reactance of the circuit can be computed from the value of L and expressed in ohms by the formula

$$\text{Reactance} = 2\pi fL.$$

The impedance of the circuit expressed in ohms is equal to the vectorial sum of the effective resistance R and the reactance. This relation is made use of in practice when it is desired to express the impedance of circuits in round numbers without reference to its component parts. Generally, however, in the practical applications of impedance measuring in maintenance work, the resistance and inductance components can be used directly to the best advantage without combining them or expressing the inductance readings in terms of reactance.

One of the most important applications of impedance measurements is the determination of the characteristic impedance of telephone circuits at the various frequencies involved in the transmission of telephone currents. Measurements of this kind, when applied to equipment circuits such as telephone repeaters, balancing networks, etc., and to the line circuits themselves, tell a great deal in regard to the efficiency of these circuits for the transmission of speech. They are very important, therefore, in checking up the installation of certain circuits in the plant and making sure that the proper impedance relations are obtained.

Fig. 17 shows the results of impedance measurements on a loaded 19 gauge cable circuit within a range of frequencies from 300 cycles to 2,300 cycles. The effective resistance values and the values of the reactance components are indicated by the curves. The inductance values are negative which means that the circuit tested had capacitive reactance throughout the range of frequencies used. When the measurements were made the distant terminal of the circuit was terminated by an impedance approximating the characteristic impedance of the circuit in order to give the effect of an infinite length of line. If the above circuit is used for 2-way telephone repeater operation it is necessary that the repeater balancing networks have impedance characteristics similar to the lines which they balance in order that the maximum repeater gain with good quality be obtained.

Measurements such as described above, in addition to giving a picture of the effective resistance and reactance of circuits at different frequencies, also provide a means for locating the irregularities and

troubles which tend to change the normal impedance characteristics. The omission of loading coils or the reversal of one loading coil winding, the installation of intermediate apparatus or of emergency cable, etc., cause impedance irregularities which are very detrimental to telephone repeater operation. The effect of these irregularities on an alternating current is to reflect some of the current back towards the sending end, this reflected current either adding to or subtracting

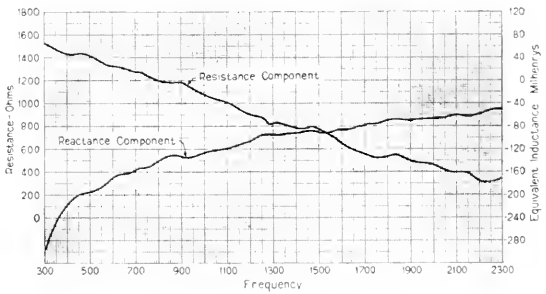


Fig. 17

from the current entering the line. This effect can be observed in impedance measurements by peaks and hollows in the effective resistance and inductance curves.

Fig. 18 shows two resistance curves of measurements made on a loaded No. 101 copper circuit (No. 12 N. B. S.), Curve *A* being for a condition where two consecutive loading coils were missing and Curve *B* for the condition after these coils were connected back in the circuit. The small irregularity in Curve *B* was due principally to the irregularity introduced by the use of a 1,500 ohm termination when making the measurement. The distance in miles from the end of the circuit at which the measurements were made to the irregularity caused by the missing loading coils is given fairly accurately by the formula:

$$\text{Distance} = \frac{V}{2(f_2 - f_1)},$$

where V is the velocity of the measuring current in miles per second for the particular type of circuit tested and $(f_2 - f_1)$ the average difference in frequencies between successive peaks of Curve *A*. For the type of circuit on which the measurements shown on Fig. 18 were

made, the velocity of propagation is approximately 54,700 miles per second. The average difference in frequencies between peaks on the curve is about 380 cycles. Applying these figures in the above formula gives the distance out to the irregularity as 70 miles. In this case the ninth and tenth loading coils were missing, which gave a very close check to the computed 70 mile figure. A great deal of use is

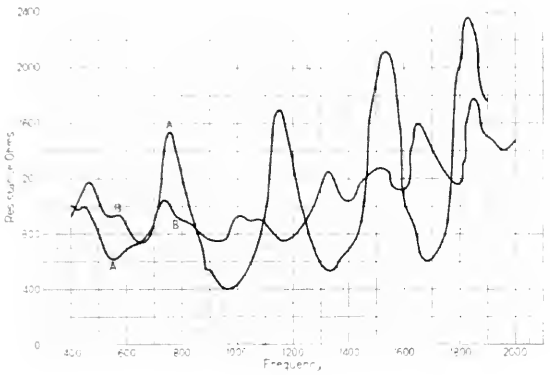


Fig. 18

made of measurements of this kind in locating troubles which affect telephone repeater operation and in directing the work of linemen in clearing these troubles.

A further use which can be made of a bridge similar to the one just described is in the location of impedance unbalance conditions which tend to increase crosstalk and noise between circuits. This is a fairly recent development and a description of it will be included in a paper to be published later.

(4) *Tests of Balance of Apparatus.* Certain types of equipment associated with telephone circuits are made up of apparatus which has to be closely balanced with respect to the various parts in order that the equipment when connected to telephone circuits will not cause unbalances in these circuits. Any unbalances introduced in this way will increase noise and crosstalk in the same manner as impedance unbalances in the line circuits themselves. Cord circuits, phantom repeating coils, composite sets, etc., are examples of the types of equipment in which unbalances in the apparatus may affect noise and

crosstalk conditions in the telephone circuits to which they are connected. The capacity bridge and the impedance bridge previously described can be used to test apparatus unbalances.

Composite sets for superposing telegraph on telephone circuits are particularly important in respect to balance and in order to provide a means for quickly checking the balance conditions in these a special form of bridge has been designed. This testing apparatus is known as the composite set bridge and is of particular advantage in that it provides for quickly testing the balance conditions of various parts

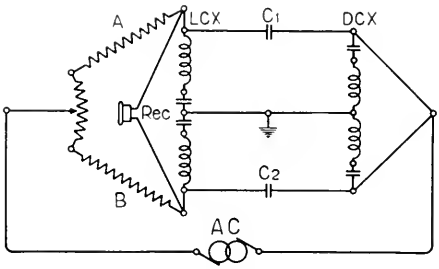


Fig. 19

of the set as well as complete sets. Tests can be made for example of the balance of the telegraph branches complete or of the condensers and coils in these branches separately. Tests can also be made of the balance of the grounded branches or of the series line condensers of the set.

To illustrate the operation of this bridge, Fig. 19 shows the arrangement for testing the balance of the series condensers in a composite set. Two arms of the bridge A and B consist of fixed resistances connected together by a slide wire resistance. The series line condensers of the composite set, C_1 and C_2, then become the other two arms of the bridge. When a source of alternating current is connected as shown, a condition of minimum tone in the receiver obtained by adjusting the position of the contactor on the slide wire indicates when the bridge is balanced. The slide wire is calibrated to read the percentage unbalance of the condensers C_1 and C_2 directly.

Crosstalk and Noise Measurements. Circuit unbalance conditions, such as described in some of the previous tests, are often very detrimental to telephone transmission in that they cause crosstalk between

circuits. Also foreign currents, induced from supply lines, produce noise which has much the same effect as inserting a transmission loss. The magnitude of noise produced in this way is dependent among other things, on the balance conditions of both the supply and telephone circuits.

The determination of the magnitude of crosstalk and of noise currents can be made by relatively simple measurements. In practice crosstalk tests, which also give an indication of the balance conditions of circuits, can be made more quickly than impedance unbalance tests, although they do not give a location directly of any troubles which may exist. The usual procedure then is to make noise and crosstalk tests on circuits, and in those cases where the measurements indicate that improvement is desirable some of the direct current or alternating current methods previously described are applied to locate the cause. The simplified circuit arrangement of the test set commonly used for measuring crosstalk between two circuits is shown in Fig. 20.

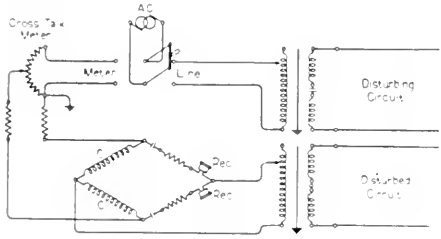


Fig. 20

An alternating current source generally of complex wave shape to simulate voice currents is connected to a switch in the set so arranged that its voltage can be impressed either on a telephone circuit known as the "Disturbing Circuit" or on a measuring shunt known as a "Crosstalk Meter." The other side of the shunt is connected through a Wheatstone bridge arrangement to a second telephone circuit known as a "Disturbed Circuit." Shielded transformers are used in the set as shown for connection to the circuits under test, these transformers being designed to give the proper impedance relations required by the different types of circuits met with in practice. The Wheatstone bridge arrangement is primarily for the purpose of allowing any noise currents which may be present in the disturbed circuit to be impressed on the observing receivers either when these are used to listen to the

cross-talk through the shunt or directly to the crosstalk from the disturbed line. Errors which might be introduced should line noise be present for only one condition of the test are in this way eliminated.

Measurements are made by first impressing the alternating current tone on the disturbing circuit and then on the meter and adjusting the shunt until the annoying effect of the tone heard in the disturbed circuit is judged to be the same as that heard on the meter. The crosstalk meter is calibrated in crosstalk units, one unit being defined as the ratio of one millionth between the current at the terminal of the disturbed circuit and the current at the terminal of the disturbing circuit, providing these currents are transmitted into like impedances and distortion of the speech sounds is not involved.

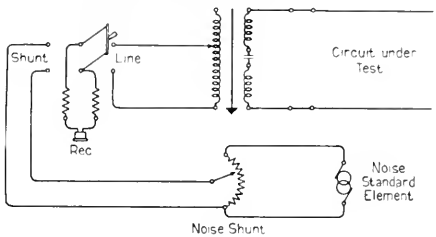


Fig. 21

Fig. 21 shows the simplified circuit of a noise measuring set arranged to measure metallic noise on a telephone circuit. As in the case of the crosstalk set, a shielded transformer is used to connect the set to the circuit under test which can be adjusted to give the proper impedance relations. With the switch thrown towards "line" the receiver is connected to the circuit under test and any noise on this circuit observed. When the switch is thrown towards "shunt" an artificial noise current produced by a vibrator is impressed on the receiver through a shunt. By alternately throwing the switch from the line under test to the shunt circuit, the shunt is adjusted until the interfering effect of the noise on the line and from the shunt are judged to be equal. The reading of the shunt which is calibrated in noise units gives a measure of the amount of noise in the circuit under test. In the commercial form of instrument used in the plant, the circuit is arranged so that both metallic noise and noise to ground can be readily measured. Where noise is present on circuits, instruments are also available for analyzing the wave shape, that is, de-

terming which frequencies making up the noise currents predominate. For both noise and crosstalk measurements, definite rules must be followed in terminating the distant ends of the circuits under test in order to reduce terminal impedance irregularities.

21-*Circuit Balance Tests.* In describing the use of the bridge for locating impedance irregularities, mention was made of the effect of such irregularities on telephone repeater operation. Since the making of impedance runs on circuits involves a considerable amount of time and expense, a simple and quick balance test, known as the 21-circuit test, was devised in which the telephone repeater is made to function as the testing set. The gain which can be obtained from a 21 or 22 type telephone repeater with good quality depends to a large extent on the degree of balance, within the frequency range involved, between the impedances of the telephone circuits and the impedances of the corresponding balancing networks. The use of this balance relation is illustrated in the simplified circuit of Fig. 22 which shows a 22 type repeater connected to make a 21-circuit balance test between the "East" line and its balancing network.

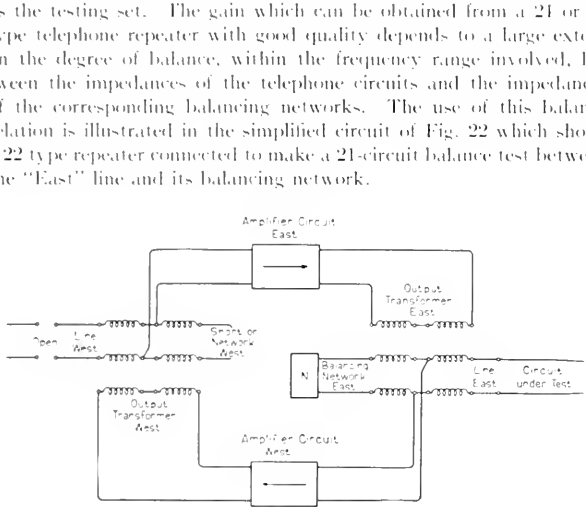


Fig. 22

The line under test and its balancing network are connected as for normal repeater operation, while the "West" line is opened. The "West" line network terminals are either shorted or the network left connected, the principle of the test being the same in either case. The 3-winding transformer, when connected for normal repeater operation, as shown for the "East" transformer in Fig. 22, simply gives a Wheatstone bridge relation, the input of the "West" amplifier being connected to the balanced points of the bridge. The proportion

of the current delivered to the transformer from the "East" amplifier which gets to the input of the "West" amplifier depends, therefore, on the degree of bridge balance furnished by the line under test and its network. When the 3-winding transformer is opened on the line side with either the network terminals shorted or with the network connected, as shown for the "West" transformer its action is the same as a repeating coil.

An internal path for currents which may produce repeater "singing" or a sustained tone, is established if the gain of the two amplifiers is just greater than the sum of the losses within the repeater circuit, that is, the losses through the transformers and any other equipment in the circuit. Theoretically, if the line and network were perfectly balanced and there were no internal unbalances in the repeater, it could not be made to sing since, due to the balance relations of the "East" 3-winding transformer, there would be infinite loss from the output of the "East" amplifier to the input of the "West" amplifier. This ideal condition is, of course, not met with in practice, since it is not practicable to design repeater circuits for perfect balance or to construct artificial networks which will exactly balance the working lines at all frequencies involved. The amplification which can be obtained in any instance without singing, then depends to a large extent on the balance between the lines and networks. In the test circuit shown in Fig. 22 the gains of the two amplifier elements are increased until singing or a sustained tone is observed and the total gain required for this gives an indication of the balance between the "East" line and its balancing network. In the same way the balance between the "West" line and its network can be determined by connecting this in the regular way to the "West" 3-winding transformer and disconnecting the "East" line. In making the tests in either direction the "poling" of the repeater circuit is reversed in order to give the lowest value of singing point which might occur under service conditions.

In practice the tests described above have become of considerable use and importance in the installation and maintenance of telephone repeaters and the circuits associated with them. Methods are available for computing the estimated singing points which circuits and equipment should give with telephone repeaters under operating conditions. These computations allow toll circuits and equipment to be engineered intelligently with respect to the gains which the repeaters may be expected to give with good quality. After installation, the 21-circuit tests furnish a means for checking computed or estimated singing points. When the estimated singing points cannot be obtained

with the 21-circuit tests, this is an indication of balance trouble which must be located either by an inspection of the circuits or balancing equipment or by resorting to impedance measurements as described previously.

Another method of determining impedance irregularities which is made use of in some of the larger offices is to measure the transmission loss through the 3-winding transformer with the lines and networks connected as for normal repeater operation. As stated previously the loss through the transformer to currents from the output of one amplifier to the input of the other gives a measure of the balance conditions of the line and network, the loss increasing as the balance becomes more perfect. By this scheme the losses through the 3-winding transformers can be measured over a range of frequencies as in line impedance measurements and a loss curve obtained which can be used to locate irregularities in the same manner as described for line impedance curves.

Transmission Efficiency Measurements. If all or a part of the tests already described were applied to the various transmission circuits in the telephone plant, most troubles which might effect speech transmission could be detected and assurance given that the circuits were properly installed. Such a procedure would be costly and impracticable and for this reason it is necessary that means be provided whereby a measurement of a circuit's efficiency for the transmission of voice currents can be quickly made.

The transmission of voice currents can be measured in terms of a standard and expressed in units in much the same manner as the transmission of any electrical currents. A telephone circuit, for example, extending between any two offices is said to have an equivalent of so many units of transmission, the number of these units depending on the electrical characteristics of the component parts of the circuit.[1] Transmission measurements, as far as volume efficiency is concerned, involve determining by means of suitable testing apparatus the number of transmission units of loss or gain which a particular circuit or piece of equipment causes. As it is desired to obtain a measure of efficiency at a frequency comparable with the combined frequencies of the voice, a frequency of 1,000 cycles for the testing current has been chosen which experience has shown gives results approximating fairly closely those obtained by using a combination of the frequencies within the voice range. Measurements can also be made at other frequencies within the voice range or at

[1] See the article in this issue, The Transmission Unit and Telephone Transmission Reference Systems, by W. H. Martin.

frequencies outside of this range where desired, for example, at ringing current frequencies or carrier current frequencies.

Efficiency tests of transmitters and receivers present a somewhat different problem and for these it has been found most convenient to make direct comparisons between the instruments under test and standard instruments.

A discussion of the application of transmission testing apparatus in maintenance work for measuring losses and gains is given below.

(1) *Measurements of Transmission Losses.* In its simplest form a transmission measuring set involves an arrangement of apparatus whereby a volume comparison can be made between voice currents transmitted over a circuit of unknown efficiency and then over a standard circuit of known efficiency.

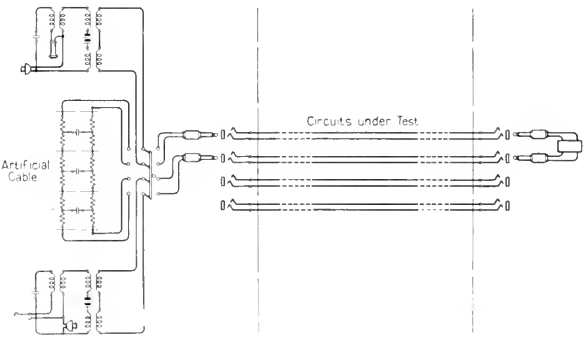


Fig. 23

Such an arrangement is illustrated in Fig. 23 in which the amount of artificial cable required to give a volume of transmission equal to that obtained over the circuit under test is a measure of the circuit's efficiency in terms of the artificial cable units. Prior to the development of the present types of transmission measuring sets the arrangement shown in Fig. 23 was used to a limited extent, principally in making measurements on important types of toll circuits and in determining fundamental transmission data such as unit equivalents, reflection losses, etc.

To meet the practical requirements of field testing work two general types of testing apparatus have been developed, one involving "ear

balance" methods and the other "visual" methods, that is, an amplifier and detector arrangement.

Fig. 24 shows the schematic circuit arrangement for an ear balance test set and Fig. 25, that for a set employing visual methods.

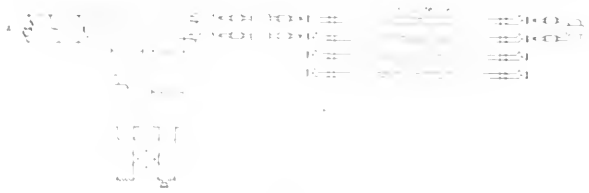


Fig. 24

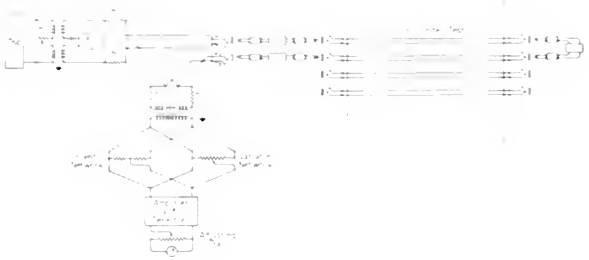


Fig. 25

A description of this apparatus and its development has been given in a paper by Best.[2] In brief, one subscriber's set of the circuit in Fig. 23 has been replaced by an oscillator while the other subscriber's set has been replaced by a receiver and resistance arrangement in the circuit of Fig. 24 and by an amplifier and detector in the circuit of Fig. 25. The artificial cable of Fig. 23 has also been replaced by distortionless resistance network standards in Figs. 24 and 25. Various resistances and coils are also provided to meet practical testing requirements such as adjusting the measuring current, and reducing reflection losses.

For field testing work, the circuits shown in Figs. 24 and 25 are mounted in compact form in portable boxes which can be readily carried from office to office or wherever required. Portable oscillators for supplying the measuring current are also provided so that com-

[2] F. H. Best, *Jour. A. I. E. E.*, Vol. XLIII, No. 2, Feb., 1924

plete testing equipment is available, by means of which, a large volume of transmission testing in central offices and private branch exchanges can be done in the most convenient manner. Tests using these instruments can be made as readily on the transmission circuits in machine switching offices of both the step by step and panel types as in manual offices. The ear balance set of Fig. 24 requires no external source of direct current power and only a three dry cell battery is required for operating the oscillator. It is, therefore, used to the best advantage in testing private branch exchange switchboards and magneto switchboards where the power necessary to operate visual types of sets is not readily available.

The visual type of set of Fig. 25 is particularly suited for testing in the larger common battery central offices since it permits measurements to be made more quickly and accurately than in the case of the

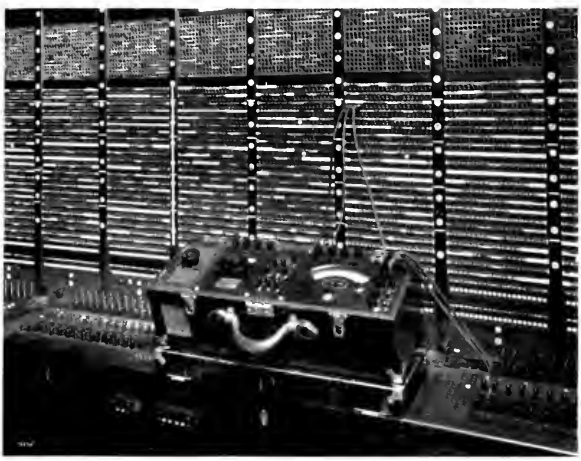


Fig. 26

ear balance sets. These larger offices also have readily available the 24-volt batteries required to operate the visual reading sets. Fig. 26 shows a picture of one of the latest types of portable visual reading measuring sets, set up ready for operation at a central office switchboard position.

In order to give a general picture of the kinds of trouble found with this transmission testing equipment the following table shows a trouble classification which is particularly useful in analyzing testing results and instigating any required remedial measures.

Classification of Troubles Found

Physical defects.	Wrong type of equipment or circuit.
Opens.	Missing equipment.
Grounds.	High resistance.
Crosses.	Low insulation.
Cut Outs.	Wrong routing.
Electrical defects.	Bridged conductors.
Incorrect wiring.	

The above classification includes all of the common types of troubles which, if not kept out of the plant, will be detrimental to service. The item of physical defects is a class of trouble which is not determined directly by transmission tests but is discovered by the maintenance forces during the course of their testing work. It represents any unsatisfactory conditions found in the circuits which, while not causing trouble at the time, may very likely do so later and should, therefore, be corrected. The next four kinds of trouble shown in the table viz: opens, grounds, crosses and cut-outs while detected by transmission tests can also be found and cleared by the every-day maintenance work without the use of transmission testing apparatus. The remaining classes of trouble listed can, it has been found, be detected and eliminated most efficiently by the use of transmission testing sets. Classifying troubles and identifying them with the important circuits in the exchange area plant such as cord circuits, operators' circuits, trunks, etc., has proved very valuable in transmission maintenance work. The results of the work when analyzed in this way are a very great aid in supervision and assist materially in keeping the plant in good condition.

The visual reading circuit of Fig. 25 is also designed in a form for permanent installation particularly for use in testing toll circuits. A picture of a typical installation of one of the latest types of sets and its associated oscillator is shown in Fig. 27.

From 40 to 50 instruments of the general type shown in the picture are now located at important toll centers throughout the country. They are constantly used to check the overall transmission efficiency

of toll circuits and in locating and clearing any transmission troubles which occur in service. Tests are made either by looping two circuits together at the distant terminals and measuring the loop loss or by testing single circuits straight-away between toll centers equipped with measuring instruments of this type.

In general, transmission testing apparatus quickly locates kinds of troubles which cannot be readily detected by other routine testing methods. Transmission tests also serve as a means for checking

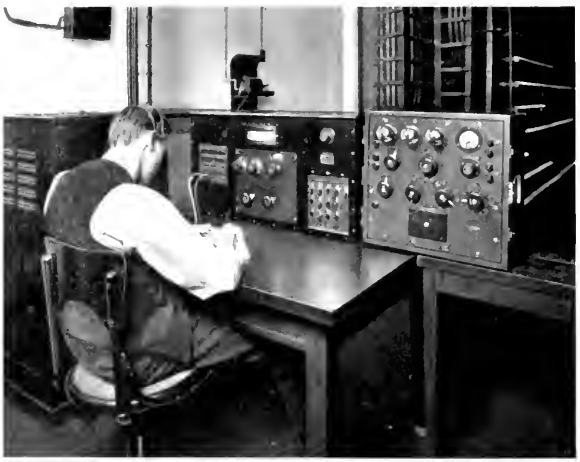


Fig. 27

general maintenance conditions and for insuring that other routine testing work is carried on in an effective manner.

As an illustration of the effect of some of the kinds of troubles which transmission tests detect, Fig. 28 shows the transmission circuit arrangement for a typical toll connection and below transmission level diagrams are given for the normal transmission condition and for conditions where common kinds of troubles are present. The level diagrams show how the normal overall transmission equivalent is increased when one or a number of transmission troubles are present in the various circuits going to make up the connection between subscribers. As indicated some troubles are more severe than others,

	Adds.	Total Equivalent
A Normal Transmission		25.0 TU
B Defective Receiving Coil	2.5 TU	27.5 "
C Defective CX Set	3.0 "	30.0 "
D Improper Repeater Adjustment	6.0 "	34.0 "
E Open Condenser in Rec. Coil	8.0 "	36.0 "
F Transmitter Down in Efficiency	10 TU	46.0 "
G Total of All Losses		56.5 TU

Fig. 28

but any of them tend to produce conditions which may be very detrimental to service. These diagrams illustrate therefore, how important it is to maintain the telephone plant so that troubles of this nature will not be present.

(2) *Measurements of Transmission Gains.* The transmission gains of amplifier circuits are measured in much the same way as transmission losses. A gain may be considered as a negative loss and is expressed in the same transmission units. In measuring the gains of amplifier circuits designed for two way operation, it is necessary to provide the proper balancing conditions in order to prevent "singing." This is done by connecting the amplifier circuit between two artificial lines of the proper impedances and balancing these lines by networks. The simplest measuring circuit now in use consists of an arrangement whereby the repeater or amplifier under test is connected between two artificial lines with balancing networks and tone is supplied by an oscillator at the terminal of one line while the terminal of the other line is equipped with a receiver and a measuring shunt calibrated in transmission units. The repeater and the shunt are then alternately cut in and out of the circuit and the shunt adjusted until equal volume of tone is observed in the receiver, for which condition the shunt reading gives the gain of the repeater.

A visual method for measuring repeater gains is provided by substituting an amplifier detector circuit for the shunt and receiver. This is essentially what is done in the transmission measuring circuit shown in Fig. 25. The type of set designed for permanent installation which employs this circuit is arranged so that amplifier gains up to about 20 TU can be measured when a repeater is connected in place of the lines under test and the necessary repeater balancing requirements taken care of. The gains of repeaters connected in toll circuits are often checked in this way when overall transmission tests are made on these circuits.

To meet practical testing requirements at the larger repeater and carrier stations where a considerable amount of gain testing work is done, a visual reading measuring set especially designed for testing amplifier gains has been developed. The measuring circuit employed in this gain testing set has been described.[3] The equipment going to make up these sets, that is, the measuring shunts, artificial lines, amplifiers, meters, etc., is mounted in compact form on standard panels which can be installed at convenient locations near repeater and carrier equipment. A panel mounted 1,000 cycle oscillator is also

[3] A. B. Clark, *Bell System Technical Journal*, Volume II, No. 1, January, 1923.

provided to supply measuring current, although other types of oscillators giving the necessary output and proper wave shape can be used if desired.

In practice, it is necessary to maintain the gains of the amplifiers in repeater and carrier circuits to fairly close limits since these amplifiers form an integral part of toll circuits. Measurements of gains are also made in connection with the 21 circuit balance tests previously described. Another important application of gain tests is to check the gain frequency characteristics of repeaters to determine that all frequencies within the voice range are being properly amplified. By varying the filament current between limits, a test of the vacuum tubes for filament activity is obtained by gain measurements.

(3) *Measurements of Transmitter and Receiver Efficiencies.* Transmitters and receivers are used in the telephone plant principally in operators' sets and subscribers' sets. In the former, the transmitters, receivers and operators' circuits are readily available to the maintenance forces and therefore can be inspected and tested in a routine manner. In the case of subscribers' sets, however, the equipment in service is not accessible and tests must be made on the instruments before installation or at times when they are removed from service. Talking tests can also be made from the instruments at the time installations are made and any particularly unsatisfactory conditions found in this way.

The difficulties incident to testing transmitters and receivers are due to the fact that in transmitters, the efficiency depends on the ability to convert sound energy into electrical energy and in receivers, the ability to convert electrical energy into sound energy. Obviously, a simple form of transmitter test and one which has until recently been generally used is to talk alternately into the transmitter under test and then into a standard transmitter and observe the difference in volumes at a receiving set. In the same manner, a simple receiver test is to listen alternately to a receiver under test and then to a standard receiver connected to a talking station. This method is slow and also of limited accuracy due to inherent changes in a speaker's voice and to the possibility of the distance of the speaker's lips from the transmitter varying. To take the place of this method transmitter and receiver testing machines have been developed which will be described in a paper to be published later.

Oscillators. Practically all alternating current testing work requires the provision of an external source of measuring current. For this purpose oscillators of various types have been developed which are designed electrically and mechanically to meet various test circuit

requirements as to wave shape, volume of current, etc. One of the earliest forms of oscillators known as the "substation howler" was made by coupling the receiver and transmitter of a subscriber's set together mechanically and taking off the alternating current generated by means of an induction coil in the howling circuit. This type of oscillator, which was subject to large variations in volume and produced a very poor wave form, has now been replaced by other and improved types.

Oscillators now in use in the field can be divided into three general classes, those employing vibrators, those employing motor generator equipment and those employing vacuum tubes. The principles of these oscillators are briefly described below by considering one commercial type in each class:

(1) *Oscillators Employing Vibrators.* Fig. 29 shows the circuit arrangement of an oscillator of this type which is designed for producing a single frequency alternating current.

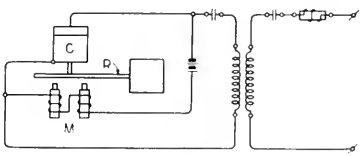


Fig. 29

The current generating element consists of a metal reed *R*, resting against the diaphragm associated with the carbon button *C*. The receiver spools *M*, are so arranged that when they are energized by the battery an attractive force is exerted upon the reed which draws it away from the carbon button. This action decreases the pressure in the carbon button with a corresponding decrease in the current from the battery, which in turn decreases the attractive force of the receiver so that the pressure of the reed against the carbon button is again increased. This cycle of change in current and pressure is repeated at the natural frequency of vibration of the reed so long as direct current flows from the battery. The alternating currents set up in this way are passed through a circuit resonant at the natural period of vibration of the reed, thereby giving a current of good wave form.

This particular form of vibrator oscillator is used principally in transmission testing work where portable "ear balance" methods

are employed. It may, however, be used for other kinds of measurements where single frequency currents of fairly good wave form are required. Other forms of vibrator oscillators are available, particularly for use in capacity and capacity unbalance tests and cross-talk and noise tests.

(2) *Oscillators Employing Motor Generator Equipment.* This type of oscillator is illustrated by ordinary ringing and trouble tone machines and the low frequency alternating currents generated by these machines are often used in testboard work. The circuit for an oscillator of this type, particularly designed for producing 1,000-cycle alternating current with good wave form, is shown in Fig. 30.

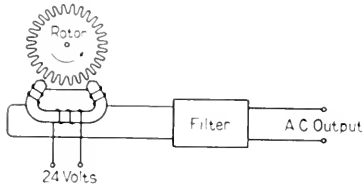


Fig. 30

In this circuit the field of an electromagnet is varied by a laminated core or rotor resembling a spur gear driven by a small 24-volt d.c. motor. The speed of the motor is automatically regulated and the electromagnet and rotor so designed that a 1,000-cycle current is generated. Harmonics which are inherent in the oscillator are eliminated by the use of a filter. This oscillator can be operated on the regular 24-volt central office battery and is compactly mounted to make it readily portable. It is particularly adaptable, therefore, for supplying the measuring current required to operate portable visual transmission measuring sets and is now generally used for this purpose in the telephone plant.

(3) *Oscillators Employing Vacuum Tubes.* Fig. 31 shows the simplified circuit arrangement of a vacuum tube oscillator.

The oscillating vacuum tube in this generator has its plate and grid inductively connected together in a tuned circuit. Closing the filament battery circuit starts this tube oscillating, the frequency of the oscillations being controlled by the inductance of the plate and grid coupling and the variable condenser C. The current thus generated is amplified by other vacuum tubes to the values which are required in the alternating current testing work. The circuit of Fig. 31 shows

only one amplifying vacuum tube, but additional amplifiers may be added to meet the requirements of particular kinds of testing work. One of the latest forms of oscillators of this type is shown in Fig. 27 set up for use with one of the permanent types of transmission measuring sets.

Vacuum tube oscillators have been developed which will generate measuring currents of any desired frequency within the range of 100 cycles to 50,000 cycles, thus covering both the voice and carrier

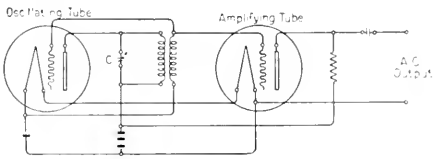
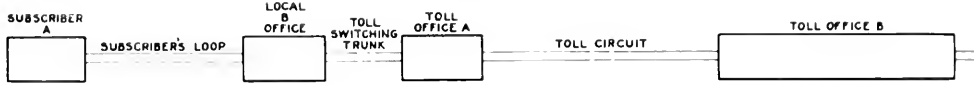


Fig. 31

range. These oscillators have become indispensable in testing and maintenance work. They are used extensively in making both single frequency transmission tests and transmission loss and gain tests within the range of frequencies mentioned above. They are also used in making line impedance and impedance unbalance tests and in determining the characteristics of telephone repeater and carrier circuits.

Specific Applications of Electrical Testing Methods. In describing the various electrical tests above, considerable has been said regarding the applications which are made of them to insure satisfactory telephone transmission. In order to give an overall picture of these applications the toll connection for which transmission level diagrams are given in Fig. 28, is shown in simplified form in Fig. 32, with various tests listed underneath the different sections of the circuit layout. Only the sections of the circuit making up the first part of the connection are shown since corresponding tests will apply to the circuits making up the second.

The tests listed in Fig. 32 are not intended to give a testing program but rather to show the various electrical testing means which are available for use in installation and maintenance work. Just what tests should be made, the frequency of making the tests and the limits to work to, to insure a high grade of transmission depend on the types of circuits and equipment involved and their relative im-



1. Transmission efficiency tests of transmitters, receivers and instrument circuit before installation

2. Talking transmission test at time of installation

3. DC current flow and resistance tests

4. Insulation tests.

5. DC capacity, voltmeter "kick" tests

Trouble Location Tests

Wheat tone bridge measurement and any of above tests required

1. Transmission tests of toll switching trunks, toll cord circuits, operators' circuits, composite sets, cord etc.

2. Tests and inspection of operators' transmitters, receiver and cord

3. DC resistance test of local coil and protector cutout.

4. Impedance balance tests of composite set

5. Insulation test of central office wiring

6. Insulation tests of cable conductors used for trunks

Trouble Location Tests

Wheat-tone bridge measurement and any of above tests required.

1. Insulation test for grounds or leakage tests

2. DC loop resistance and to set up balance tests

3. Overall transmission efficiency tests

4. Overall noise and cross talk test

5. Composite unbalance to determine installation of shielded toll cable

Trouble Location Tests

1. Wheat tone bridge measurements for opens, shorts, and ground.

2. Impedance measurements for unbalances affecting telephone repeater operation not evident by DC tests and inspections

3. Impedance unbalance measurements for location of unbalance causing noise and crosstalk.

4. Current flow continuity tests to check open wire pin positions and transpositions.

5. Use of aerial test set by lineman to obtain exact location of troubles which have been approximately located by above test

Fig. 32

1. Transmission efficiency tests of circuits and equipment

2. Inspection and test of carrier telephone transmitters, receivers and sets.

3. Impedance [illegible] inspection sets

4. Insulation test of entire installation

5. DC resistance tests of local coils and protector gaps

Telephone Repeater Tests On

1. VU Capacity test of entire station circuits including repeater after installation

2. Direct current and [illegible]

3. Circuit balance tests

4. Balance of [illegible] Volume units

5. Transmission [illegible] operating steps

6. Transmission [illegible] tests

7. Transmission [illegible]

Trouble Location Is

Same tests as above. In addition characteristic impedance tests repeaters and networks [illegible] necessary.

portance in the system as a whole. These matters are covered in routine instructions which are developed by experience and which take into account local conditions and service requirements.

In conclusion, it may be stated that telephone systems in their present development have at their disposal means for carrying on an adequate transmission maintenance program in an economical manner. Furthermore, studies and trials of new methods are continually being carried on with a view to obtaining further improvements and increased economies in transmission testing and maintenance work.

A Generalization of the Reciprocal Theorem

By JOHN R. CARSON

THE Reciprocal Theorem, an interesting and extremely important relation of wide applicability, which was discovered by Lord Rayleigh, is stated by him in the language of electric circuit theory as follows:

"Let there be two circuits of insulated wire A and B, and in their neighborhood any combination of wire circuits or solid conductors in communication with condensers. A periodic electromotive force in the circuit A will give rise to the same current in B as would be excited in A if the electromotive force operated in B."[1]

Before proceeding with the generalization which is the subject of this paper, Rayleigh's theorem, in the following modified form, will first be stated and proved:

I. *Let a set of electromotive forces* $V_1' \ldots V_n'$, *all of the same frequency, acting in the n branches of an invariable network, produce a current distribution* $I_1' \ldots I_n'$, *and let a second set of electromotive forces* $V_1'' \ldots V_n''$ *of the same frequency produce a second current distribution* $I_1'' \ldots I_n''$. *Then*

$$\sum_1^n V_i' I_i'' = \sum_1^n V_i'' I_i' \tag{1}$$

To prove this theorem we start with the equations of the network

$$\sum_1^n Z_{ij} I_i = V_j, \qquad j = 1, 2, \ldots n, \tag{2}$$

and observe that, provided the network is invariable, contains no internal source of energy or unilateral device, and provided that the applied electromotive forces $V_1 \ldots V_n$ are all of the same frequency, say ω, the mutual impedances satisfy the reciprocal relations $Z_{ij} = Z_{ji}$. Consequently if (2) is solved for the currents, we get

$$I_j = \sum_1^n \lambda_{ij} V_i, \qquad j = 1, 2 \ldots n, \tag{3}$$

and the coefficients also obey the reciprocal relations $\lambda_{ij} = \lambda_{ji}$.

Now consider two independent and arbitrary sets of equi-periodic applied electromotive forces, $V_1' \ldots V_n'$ and $V_1'' \ldots V_n''$; then

[1] Rayleigh, Theory of Sound, Vol. I, p. 155.

in accordance with (3), the corresponding distributions of network currents $I_1' \ldots I_n'$ and $I_1'' \ldots I_n''$ are given by

$$I_j' = \sum_{k=1}^{n} A_{jk} V_k', \qquad j = 1, 2 \ldots n, \tag{4}$$

$$I_j'' = \sum_{k=1}^{n} A_{jk} V_k''. \tag{5}$$

Now form the product sum $\sum V_j'' I_j'$; by means of (4) it is easy to show that, since $A_{jk} = A_{kj}$,

$$\sum_{j=1}^{n} V_j'' I_j' = \sum_{j=1}^{n} \sum_{k=1}^{n} A_{jk} (V_j' V_k'' + V_j'' V_j') - \sum A_{jj} V_j' V_j''.$$

Since this is symmetrical in the two sets of applied forces $V_1' \ldots V_n'$ and $V_1'' \ldots V_n''$, it follows at once that

$$\sum V_j'' I_j' = \sum V_j' I_j'',$$

which proves the theorem.

Now if we analyze the foregoing proof it is seen to depend on the assumption, first that the network can be described in terms of a set of simultaneous equations with constant coefficients, and secondly on the reciprocal relation in the coefficients, $Z_{jk} = Z_{kj}$. In other words, it is assumed that the currents flow in linear, invariable circuits, and that the system is what is called quasi-stationary.[2] What this means is that the network may be treated as a dynamical system defined by n coordinates, the n currents $I_1 \ldots I_n$ being the velocities of the n coordinates. More precisely stated, the underlying assumption is that the magnetic energy, the electric energy, and the dissipation function can be expressed as homogeneous quadratic functions of the following form

$$T = \tfrac{1}{2} \sum \sum L_{jk} I_j I_k,$$

$$W = \tfrac{1}{2} \sum \sum S_{jk} Q_j Q_k, \qquad I_j = d\ Q_j / dt,$$

and

$$D = \tfrac{1}{2} \sum \sum R_{jk} I_j I_k,$$

where the coefficients L_{jk}, S_{jk}, R_{jk} are constants. Subject to these assumptions, which, it may be remarked, underlie the whole of electric circuit theory, the direct application of Lagrange's equations to the quadratic functions T, W, D leads at once to the circuit equations (1) and the reciprocal relation $Z_{jk} = Z_{kj}$. This is merely a very brief outline of Maxwell's dynamical theory of quasi-stationary systems or networks.

[2] See Theorie der Electrizitat, Abraham u. Foppl, Vol. I, p. 254.

Now in view of the foregoing assumptions and restrictions which underlie all the proofs of the Reciprocal Theorem, known to the writer, it is by no means obvious that the theorem is valid when we have to do with currents in continuous media as well as in linear circuits, and when, furthermore we have to take account of radiation phenomena.[3] The proof or disproof of the theorem in the electromagnetic case is, however, extremely important. The writer therefore, offers the following generalized Reciprocal Theorem, subject to the restriction noted below.

II. *Let a distribution of impressed periodic electric intensity* $\boldsymbol{F}'=\boldsymbol{F}'(x, y, z)$ *produce a corresponding distribution of current intensity* $\boldsymbol{u}'=\boldsymbol{u}'(x, y, z)$, *and let a second distribution of equi-periodic impressed electric intensity* $\boldsymbol{F}''=\boldsymbol{F}''(x, y, z)$ *produce a second distribution of current intensity* $\boldsymbol{u}''=\boldsymbol{u}''(x, y, z)$, *then*

$$\int(\boldsymbol{F}'\cdot\boldsymbol{u}'')\mathrm{d}v=\int(\boldsymbol{F}''\cdot\boldsymbol{u}')\mathrm{d}v, \tag{6}$$

the volume integration being extended over all conducting and dielectric media. $\boldsymbol{F}$ and $\boldsymbol{u}$ are vectors and the expression $\boldsymbol{F}\cdot\boldsymbol{u}$ denotes the scalar product of the two vectors.

The only serious restriction on the generality of this theorem, as proved below, is that magnetic matter is excluded; in other words it is assumed that all conducting and dielectric media in the field have unit permeability. This restriction is theoretically to be regretted, but is not of serious consequence in important practical applications.

Proof of Generalized Reciprocal Theorem[4]

In order to prove the generalized theorem stated above it is necessary to discard the special assumption of quasi-stationary systems underlying Rayleigh's theorem, and start with the fundamental equations of electromagnetic theory. These may be formulated as follows:

$$\operatorname{div}\boldsymbol{B}=0,$$

$$\operatorname{div}\boldsymbol{E}=4\pi\rho,$$

$$\operatorname{curl}\boldsymbol{E}=-\frac{1}{c}\frac{\partial}{\partial t}\boldsymbol{B},$$

$$\operatorname{curl}\boldsymbol{B}=4\pi\boldsymbol{u}+\frac{1}{c}\frac{\partial}{\partial t}\boldsymbol{E},$$

where c is the velocity of light.

[3] The theory of quasi-stationary systems expressly excludes radiation.

[4] In the following proof it is necessary to assume a knowledge on the part of the reader of the elements of vector analysis; the notation is that employed by Abraham

It will be noted that there are only two field vectors, $\boldsymbol{E}$ and $\boldsymbol{B}$, instead of the usual four vectors $\boldsymbol{E}$, $\boldsymbol{D}$, $\boldsymbol{B}$, $\boldsymbol{H}$, where $\boldsymbol{D}=k\boldsymbol{E}$ and $\boldsymbol{B}=\mu\boldsymbol{H}$, and that the constants of the medium k and μ do not explicitly appear. This formal simplification is effected by taking as the current density

$$\boldsymbol{u}=\overline{\boldsymbol{u}}+\frac{1}{c}\frac{\partial \boldsymbol{P}}{\partial t}+\operatorname{curl}\boldsymbol{M}$$

where $\overline{\boldsymbol{u}}$ is the conduction current density, $\boldsymbol{P}$ is the polarization, defined as

$$\boldsymbol{P}=\frac{k-1}{4\pi}\boldsymbol{E},$$

and $\boldsymbol{M}$ is defined as

$$\boldsymbol{M}=\frac{1}{4\pi}\frac{\mu-1}{\mu}\boldsymbol{B}.$$

The equation of continuity

$$\operatorname{div}\boldsymbol{u}=-\frac{1}{c}\frac{\partial}{\partial t}\rho$$

then determines the charge density ρ.

The advantage of this formulation is that $\boldsymbol{E}$ and $\boldsymbol{B}$ can now be expressed in terms of the retarded scalar and vector potentials Φ and $\boldsymbol{A}$, as follows:

$$\boldsymbol{E}=-\frac{1}{c}\frac{\partial}{\partial t}\boldsymbol{A}-\nabla\phi,$$

$$\boldsymbol{B}=\operatorname{curl}\boldsymbol{A},$$

where

$$\Phi=\int\frac{\rho(t-r/c)}{r}\,dv,$$

$$\boldsymbol{A}=\int\frac{\boldsymbol{u}(t-r/c)}{r}\,dv.$$

The notation $\rho(t-r/c)$ and $\boldsymbol{u}(t-r/c)$ indicates that ρ and $\boldsymbol{u}$ are taken not at time t but at time $t-r/c$ in evaluating the integrals. It will be observed that with ρ and $\boldsymbol{u}$ defined as above, all effects are transmitted with the velocity of light, independently of the characteristics of the medium, a point of view in accordance with the modern development of electromagnetic theory.

In the application of the preceding equations to our problem, it will be assumed that $\boldsymbol{M}$ is everywhere zero, so that

$$\boldsymbol{u} = \overline{\boldsymbol{u}} + \frac{1}{c}\frac{\partial \boldsymbol{P}}{\partial t}.$$

It will be assumed further that $\overline{\boldsymbol{u}} = \sigma \boldsymbol{E}$ and, since $\boldsymbol{P} = \frac{k-1}{4\pi}\boldsymbol{E}$,

$$\boldsymbol{u} = \left(\sigma + \frac{k-1}{4\pi}\frac{1}{c}\frac{\partial}{\partial t}\right)\boldsymbol{E}$$

and is therefore a linear function of $\boldsymbol{E}$. σ and k are in general point functions of the medium. The reason for setting $\boldsymbol{M} = 0$, is that it appears essential to the following proof that $\boldsymbol{u}$ shall be linear in $\boldsymbol{E}$; that is, that the current density at any point be proportional to the electric intensity.[3]

With the foregoing very brief review of the fundamental equations, we are now prepared to prove the generalized reciprocal theorem. Assuming a periodic steady state, so that $\partial/\partial t = i\omega$, we start with the vector equation

$$\boldsymbol{E} = \boldsymbol{F} - \frac{i\omega}{c}\boldsymbol{A} - \nabla\phi \tag{7}$$

where

$$\boldsymbol{A} = \int \frac{1}{r}\exp\left(-\frac{i\omega}{c}r\right)\boldsymbol{u}\,dv,$$

$$\Phi = \int \frac{1}{r}\exp\left(-\frac{i\omega}{c}r\right)\rho\,dv.$$

Here F is the *impressed intensity*: that is, *the electric intensity which is not due to the currents and charges of the system itself*. Also by virtue of the assumption $\boldsymbol{M} = 0$,

$$\boldsymbol{u} = \left(\sigma + \frac{k-1}{4\pi}\frac{i\omega}{c}\right)\boldsymbol{E} = \lambda \boldsymbol{E},$$

whence (7) can be written as

$$\frac{1}{\lambda}\boldsymbol{u} + \frac{i\omega}{c}\int \frac{1}{r}\exp\left(-\frac{i\omega}{c}r\right)\boldsymbol{u}\,dv = \boldsymbol{G}, \tag{8}$$

where $\boldsymbol{G} = \boldsymbol{F} - \nabla\Phi$.

[3] The question as to whether the generalized theorem itself, and not merely the foregoing proof, is restricted in general to the case where $\boldsymbol{M}$ is everywhere zero has not as yet received a conclusive answer. There are reasons, however, which cannot be fully entered into here, which make it appear probable that the theorem itself is in general restricted to the case where the current density contributing to the retarded vector potential is linear in the electric intensity and the two vectors are parallel. Subject to the hypothesis and assumptions of quasi-stationary systems, however, the restriction $\boldsymbol{M} = 0$ is not necessary. The writer hopes to deal with these questions in a future paper.

Equation (8) is a vector integral equation[6] in $\boldsymbol{u}$. The nucleus or kernel of the equation, $\exp\left(\frac{i\omega}{c}r\right)/r$, is symmetrical with respect to any two points $(x_1y_1z_1)$ and $(x_2y_2z_2)$, the distance between which is r. By virtue of this symmetry the following reciprocal relation is easily established.[7]

If $\boldsymbol{u}' = \boldsymbol{u}'(x,y,z)$ *is a function satisfying equation* (8) *when* $\boldsymbol{G} = \boldsymbol{G}' = \boldsymbol{G}'(x, y, z)$ *and* $\boldsymbol{u}'' = \boldsymbol{u}''(x, y, z)$ *a second function satisfying* (8) *when* $\boldsymbol{G} = \boldsymbol{G}'' = \boldsymbol{G}''(x, y, z)$, *then*

$$\int (\boldsymbol{u}' \cdot \boldsymbol{G}'')\,d\mathrm{v} = \int (\boldsymbol{u}'' \cdot \boldsymbol{G}')\,d\mathrm{v}. \tag{9}$$

Consequently since $\boldsymbol{G} = \boldsymbol{F} - \nabla\Phi$

$$\int (\boldsymbol{u}' \cdot \boldsymbol{F}'')\,d\mathrm{v} - \int (\boldsymbol{u}'' \cdot \boldsymbol{F}')\,d\mathrm{v} = \int \left\{ (\boldsymbol{u}' \cdot \nabla\Phi'') - (\boldsymbol{u}'' \cdot \nabla\Phi') \right\} d\mathrm{v}. \tag{10}$$

The proof of the theorem is now reduced to showing that

$$\int \left\{ (\boldsymbol{u}' \cdot \nabla\Phi'') - (\boldsymbol{u}'' \cdot \nabla\Phi') \right\} d\mathrm{v} = 0.$$

Now integrating by parts

$$\int (\boldsymbol{u}' \cdot \nabla\Phi'')\,d\mathrm{v} = -\int \Phi'' \operatorname{div} \boldsymbol{u}'\,d\mathrm{v},$$

$$= \frac{i\omega}{c}\int \Phi''\rho'\,d\mathrm{v},$$

since, from the equations of continuity, $\operatorname{div} \boldsymbol{u} = -\frac{i\omega}{c}\rho$. But from the fundamental field equations:

$$1\pi\rho' = -\nabla^2\Phi' + \left(\frac{i\omega}{c}\right)^2 \Phi'$$

whence

$$\int \left\{ (\boldsymbol{u}' \cdot \nabla\Phi'') - (\boldsymbol{u}'' \cdot \nabla\Phi') \right\} d\mathrm{v} = \frac{1}{1\pi}\left(\frac{i\omega}{c}\right)^2 \int \left\{ \Phi'\nabla^2\Phi'' - \Phi''\nabla^2\Phi' \right\} d\mathrm{v},$$

and by Greens Theorem, the right hand volume integral is equal to the surface integral

$$\frac{1}{1\pi}\left(\frac{i\omega}{c}\right)^2 \int \left\{ \Phi' \frac{\partial}{\partial n}\Phi'' - \Phi'' \frac{\partial}{\partial n}\Phi' \right\} dS,$$

the surface being any surface which totally encloses the volume, and $\frac{\partial}{\partial n}$ denoting differentiation along the normal to the surface.

[6] The formulation of the electromagnetic field equations in this form is of considerable importance. The integral equation furnishes a basis for developing electric circuit theory from the fundamental field equations. In addition it leads to the solution of problems in wave propagation which can not be directly solved from the wave equation itself.

[7] Perhaps the easiest way to prove this proposition is to regard the integral equation as the limit of a set of simultaneous equations, a point of view which forms the basis of Fredholm's researches on integral equations.

Now if the surface be taken as a sphere of radius R, centered at or near the system, it is easily shown that if R is taken sufficiently large

$$\frac{\partial}{\partial n}\Phi' = \frac{\partial}{\partial R}\Phi' = -\frac{i\omega}{c}\Phi',$$

$$\frac{\partial}{\partial n}\Phi'' = -\frac{i\omega}{c}\Phi'',$$

and the surface integral vanishes. Consequently we have established the *generalized reciprocal theorem*

$$\int \boldsymbol{u}' \cdot \boldsymbol{F}'' \, dv = \int \boldsymbol{u}'' \cdot \boldsymbol{F}' \, dv.$$

The Reciprocal Theorem I has long been employed in electric circuit theory, and has proved extremely useful. As an example of the practical utility of the generalized theorem II it may be remarked that it enables us to deduce the transmitting properties of an antenna system from its receiving properties. The latter may sometimes be approximately deduced quite simply, as in the case of the wave antenna, whereas a direct theoretical determination of the former presents enormous difficulties.

The Transmission Unit and Telephone Transmission Reference Systems[1]

By W. H. MARTIN

Synopsis: Consideration is given to the method of determining and expressing the transmission efficiencies of telephone circuits and apparatus, and of the desirable qualifications for a unit in which to express these efficiencies. The "transmission unit" described in this paper has been selected as being much more suitable for this purpose under present conditions than the "mile of standard cable" which has been generally used in the past.

THE "mile of standard cable" has been used in telephone engineering in this country for over twenty years, and during that time has been adopted in other countries, as the unit for expressing the transmission efficiency of telephone circuits and apparatus. In the present state of the telephone art, this unit has been found, however, to be not entirely suitable and it has recently been replaced in the Bell System by another unit which for the present, at least, has been called simply the "transmission unit." Before considering the reasons for such a fundamental change and the relative merits of the two units, it may be well to review briefly the general method of determining the efficiency of such circuits and the apparatus associated with them.

The function of a telephone circuit is to reproduce at one terminal the speech sounds which are impressed upon it at the other terminal. The input and output of the circuit are in the form of sound and its efficiency as a transmission system may be expressed as the ratio of the sound power output to the sound power input. For commercial circuits, this ratio may be of the order of 0.01 to 0.001.

In the operation of the system, the sound power input is converted by the transmitter into electrical power, which is transmitted over the line to the receiver and there reconverted into sound power. The effect of inserting a section of line or piece of apparatus or of making any change in the circuit can be determined in terms of the variation which it produces in the ratio of the sound power output to the sound power input, or, if this latter is kept constant, in terms of the ratio of sound power output after the change to that obtained before the change was made. It should be noted particularly that the change in the output power of the system is the real measure of the effect of any part of the circuit on the efficiency of the system and that the ratio of the power leaving any part to that entering it is not necessarily the measure of this effect. For example, a pure

[1] Reprinted from the *Journ. A. I. E. E.*, for June, 1924.

reactance placed in series between the transmitter and the line, may change the power delivered to the line by the transmitter and hence the output of the receiver, the magnitude and direction of the change being determined by the impedance relations at the point of insertion. The ratio of the power leaving the reactance to that entering it is, of course, unity, as no power is dissipated in a pure reactance. In other words, the transmission efficiency of any part of a circuit cannot be considered solely from the standpoint of the ratio of output to input power for that part, or the power dissipated in that part, but must be defined in terms of its effect on the ratio of output to input power for the whole system.

By determining the effect of separately inserting the many pieces of apparatus that may form parts of typical telephone circuits, an index can be established for each of these parts of its effect on the efficiency of the circuit for the conditions of which the circuit tested is typical. Similarly, the power dissipated in unit lengths of the various types of line can be determined by noting the change in power output of the receiver caused by increasing any line by a unit length. Such indices of the transmission efficiencies of the various parts of a circuit obviously have many applications in designing and engineering telephone circuits. These indices could be taken as the ratios expressing the change in the output power of the system. This, however, has certain disadvantages. For example, the combined effect of a number of parts would then be expressed as a product of a number of ratios. Likewise, for the case of a number of parts n of the same type in series, such as a line n miles in length, the effect would be expressed as the ratio for one part or one mile of the line, raised to the nth power. In many cases, these ratios and the powers to which they would need to be raised would be such as to make their handling cumbersome. If, however, these indices are expressed in terms of a logarithmic function of a ratio selected as a unit, the sum of any number of such indices for the parts of a circuit is the corresponding index for the power ratio giving the effect of the combination of these parts.

The "mile of standard cable" is such a logarithmic function of a power ratio. The new unit also meets this important requirement.

Definition of the Transmission Unit

The "transmission unit" (abbreviated *TU*) has been chosen so that two amounts of power differ by one transmission unit when they are in the ratio of $10^{0.1}$ and any two amounts of power differ by N units when they are in the ratio of $10^{N \cdot 0.1}$. The number of trans-

mission units corresponding to the ratio of any two powers P_1 and P_2, is then the common logarithm (logarithm to the base 10) of the ratio $P_1 P_2$, divided by 0.1. This may be written $N = 10 \log_{10} P_1 P_2$. Since N is a logarithmic function of the power ratio, any two numbers of units, N_1 and N_2, corresponding respectively to two ratios, $P_a P_b$

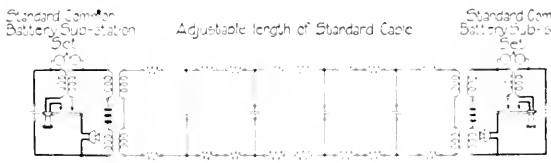


Fig. 1

and $P_c P_d$, may be added and the result $N_1 + N_2$, will correspond to the product of the ratios, $P_a/P_b \times P_c/P_d$.

From the above it is seen that the measure in transmission units of the ratio of two amounts of power P_1 and P_2 is N, where

$$N = \frac{\log \frac{P_1}{P_2}}{\log 10^{0.1}}.$$

In other words, the transmission unit is a logarithmic measure of power ratio and is numerically equal to $\log 10^{0.1}$.

The reasons for the selection of this unit and the method of applying it, can probably be best brought out by a consideration of the practise which has been followed in determining and expressing the efficiencies of telephone circuits and apparatus in terms of "miles of standard cable."

Standard Reference Circuit

Fig. 1 shows what has been designated the "standard reference circuit." It consists of two common battery telephone sets of the type standard in the Bell System at the time this circuit was adopted, connected through repeating coils or transformers to a variable length of "standard cable." This cable is an artificial line having a resistance of 88 ohms and a capacity of 0.054 microfarad per loop mile which is representative of the type of telephone cable then generally used in this country.

For a given loudness of speech sounds entering the transmitter at one end of the circuit, the loudness of the reproduced sounds given out by the receiver at the other end can be varied by changing the amount of standard cable in the circuit. Also, the amount of cable in the circuit can be used to express the ratio of the power of the reproduced sounds to that of the impressed sounds. Due to the dissipation of electrical power in the cable, this ratio and consequently the loudness of the reproduced sounds become less as the amount of cable is increased and greater as the length of cable is decreased.

This circuit then became the measuring or reference system for engineering the telephone plant and the "mile of standard cable" became the unit in which the measurements were expressed. This circuit was used to set the service standards in designing and laying out the telephone plant. Thus, the reproduction obtained over this circuit with a length of cable of about twenty miles was found suitable and practicable for local exchange, that is, intra-city service, and that corresponding to about thirty miles for toll or intercity service.

Any telephone circuit was rated by its comparison with the standard circuit. This comparison was on the basis of a speaker talking alternately over the circuit to be measured and the standard circuit and a listener switching similarly at the receiving ends, the amount of cable in the standard circuit being adjusted until the listener judged the volume of the sounds reproduced by the two systems to be equal. The number of miles of cable in the standard circuit was then used as the "transmission equivalent" of the circuit under test. The effect of any change in the circuit under test on the efficiency of that circuit could then be measured by determining the variation in the amount of standard cable required to make the sounds reproduced by the two systems again equal and the number of miles of standard cable required to compensate for this change was used as the index of this effect. In this way the relative efficiencies of two transmitters or receivers could be determined. Likewise, the power dissipation per unit length or the attenuation, of the trunk in the circuit under test could be equated to miles of standard cable. Since in each case, the standard cable is used to adjust the volume of the reproduced sound, "the mile of standard cable" corresponds to the ratio of two amounts of sound power, or as this change in sound power is produced by changing the power delivered to the telephone receiver, to a ratio of two amounts of electrical power.

If the addition of a mile of standard cable to a long trunk of the standard circuit causes the power reaching the end of the trunk to decrease by a ratio r, then the insertion of two miles will decrease

the received power by a ratio of r^2 of that obtained before the two miles were inserted. A number of miles of cable, n, inserted will reduce the received power to a ratio r^n. Thus the power ratio corresponding to any given number of miles of cable is an exponential function of the ratio corresponding to one mile, the exponent being the length in miles. The length in miles is, therefore, a logarithmic function of the power ratio.

In an infinite length of uniform line having resistance, inductance, capacity and conductance of R, L, C and G per unit length, the attenuation a per unit length, of a current of frequency f flowing along the line can be shown to be equal to the real part of the expression

$$a+j\,b=\sqrt{(R+j\,2\,\pi\,f\,L)\,(G+j\,2\,\pi\,f\,C)}.$$

For the standard cable line, since L and G are zero

$$a=\sqrt{\pi\,f\,R\,C},$$

and since $R=88$ ohms and $C=0.054$ microfarad per mile the current attenuation per mile of standard cable is

$$a=0.00386\sqrt{f}.$$

If I_1' and I_2' are the currents, respectively, at the beginning and end of a mile of line, then

$$I_1'\ I_2'=e^a \text{ or } a=\log_e I_1'\ I_2'.$$

Similarly if I_1 and I_2 are the currents, at points 1 and 2, respectively, at the beginning and end of a section of l miles

$$I_1\ I_2=e^{la} \text{ and } l\,a=\log_e I_1\ I_2.$$

For this case, the effect of inserting the section of l miles into the line on the current at point 2, or at any point beyond 2, is that the currents at the point before and after the insertion are in the same ratio as $I_1\ I_2$. Furthermore, since the impedance of the line looking toward the receiving end is the same at points 1 and 2 (and at any other points), then the ratio of the powers at the two points is equal to the square of the current ratio.

Thus the power attenuation is represented by

$$P_1\ P_2=(I_1\ I_2)^2=e^{2la}.$$

Similarly for a line, terminated in a fixed impedance which may be different from the characteristic impedance of the line, the ratio of the powers received before and after a change in the length of the

line is equal to the square of the ratio of the corresponding currents. On the basis of this relation, and because it is in general more convenient to measure or compute currents than powers, the current ratio has often been used in determining the equivalent of any piece of apparatus or line in terms of standard cable. It should be noted, however, that such a current ratio can be properly used as an index of the transmission efficiency of a part of a circuit only when it is equal to the square root of the ratio of the corresponding powers. Also, of course, the voltage ratio can be similarly used when it meets the same requirement.

Limitations in Use of Standard Cable Unit

As has been shown above, the attenuation, either of current or power, corresponding to the mile of standard cable is directly proportional to the square root of the frequency of the current under consideration. This means that the standard cable mile corresponds not only to a certain volume change in the reproduced speech sounds, but also to a distortion change. For comparisons between the standard cable circuit and commercial circuits with talking tests and as long as most of the commercial circuits had distortion comparable to that of standard cable, this two-fold effect of standard cable was desirable. At present, however, many types of circuits are being used which have much less distortion than standard cable. Also, the use of voice testing has been largely given up in the plant and it is now the general practise to determine the efficiency of circuits and apparatus on the basis of measurements and computations for single-frequency currents, a correlation having been established between these latter results and those of voice tests. These factors have made it desirable to have a unit for expressing transmission efficiencies which is distortionless, that is, is not a function of frequency.

Qualifications of a New Unit

The consideration of a new unit for measuring transmission efficiency brought out the following desirable qualifications:

1. *Logarithmic in Character.* Some of the reasons for this have already been discussed. In addition, the application of such a unit in measurements of sound make a logarithmic unit desirable, since the sensation of loudness in the ear is a logarithmic function of the energy of the sound.

2. *Distortionless.* The advantages of a unit which is independent of frequency have been referred to above. In expressing the effi-

ciency of the transmission of the high frequencies involved in carrier and radio circuits, such a unit is particularly desirable.

3. *Based on Power Ratio.* This is desirable because the power ratio is the real measure of transmission efficiency. As pointed out above, the current ratio can be used only when it is equal to the square root of the power ratio. Having the unit based on a power ratio does not, of course, require that measurements or computations be made on a power basis.

In considering the conversions between sound and electrical energy, it is obviously advantageous to have a unit based directly on a power ratio.

4. *Based on Some Simple Relation.* This is desirable in connection with the matter of getting a unit which may be widely used and may find applications in several fields.

5. *Approximately Equal in Effect on Volume to a "Mile of Standard Cable."* One reason for this is the practical one of avoiding material changes in the conceptions which have been built up regarding the magnitude of such things as transmission service standards. Also, the sound power changes which can be detected by an ear are of the order of that corresponding to a mile of standard cable. In measuring telephone lines and apparatus with single-frequency currents, it has been found that an accuracy of about one-tenth of a mile can be obtained readily and is sufficient practically.

6. *Convenient for Computations.* This refers to the matter of changing from computed or measured current or power ratios to transmission units or vice versa.

Properties of the Transmission Unit

A consideration of the above qualifications and of the various units suggested, led to the adoption of the power ratio of $10^{0.1}$ as the most suitable ratio on which to base the unit of transmission efficiency. The transmission unit is logarithmic, distortionless, is based on a power ratio and its relation to that ratio is a simple one. Its effect on the transmission of telephonic power corresponding to speech sounds is about 6 per cent less than that of one mile of standard cable. Regarding its use in computations, it has the advantage that the number of units corresponding to any power ratio, or current ratio, can be determined from a table of common logarithms.

For a power ratio of 2, the logarithm is 0.301 and the corresponding number of units is, therefore, this logarithm multiplied by 10, which is 3.01 *T U*. For a power ratio of 0.5, the logarithm is $9.699-10=-0.301$ and the number of units is -3.01 *T U*. A power ratio of 2

represents a gain of 3.01 units, and a power ratio of 0.5 corresponds to a loss of 3.01 units. If the above ratios were for current, the logarithms would be multiplied by 20. Thus a current ratio of 2 corresponds to a gain of 6.02 units and a current ratio of 0.5 corresponds to a loss of 6.02 units.

It will be noted that the $T\ U$ is based on the same ratio $10^{0.1}$ as the series of preferred numbers which has been used in some European countries and has been proposed here as the basis for size standardization in manufactured articles.[2] In common with this series, the $T\ U$ has the advantage that many of the whole numbers of units correspond approximately to easily remembered ratios as shown in the following table.

APPROXIMATE POWER RATIO

Transmission Units	For Losses		For Gains
	Fractional	Decimal	Decimal
1	4/5	0.8	1.25
2	2/3	0.63	1.6
3	1/2	0.5	2.
4	2/5	0.4	2.5
5	1/3	0.32	3.2
6	1/4	0.25	4
7	1/5	0.2	5.
8	1/6	0.16	6.
9	1/8	0.125	8
10	1/10	0.1	10.
20	1/100	0.01	100.
30	1/1000	0.001	1000.

It will be seen that the ratio for a gain of a given number of $T\ U$ is the reciprocal of the ratio for a loss of the same number of units. Also for an increase of 3 in the number of units, the loss ratio is approximately halved and the gain ratio doubled. If the approximate loss ratios corresponding to 1, 2 and 3 units are remembered, the others can be easily obtained.

From this consideration of the properties of the transmission unit, it is evident that there is much to commend its use in telephone transmission work. Furthermore, since its advantages are not peculiar to this work, such a unit may find applications in other fields. It is now being used in some of the work on sound.

[2] Size Standardization by Preferred Numbers, C. F. Hirshfeld and C. H. Berry, *Mechanical Engineering*, December, 1922.

New Telephone Transmission Reference System

With the standardization of the distortionless unit of transmission it is desirable also to adopt for a transmission reference system a telephone circuit which will be distortionless from sound input to the transmitter to sound output from the receiver. This system will consist of three elements, a transmitter, a line and a receiver. Each will be designed to be practically distortionless and the operation of each will be capable of being defined in definite physical units so that it can be reproduced from these physical values. Thus the transmitter element will be specified in terms of the ratio, over the frequency range, of the electrical power output to the sound power input, this ratio being expressed in transmission units. The receiver element will be specified likewise in terms of the ratio of sound power output to electrical power input. The output impedance of the transmitter and the input impedance of the receiver elements will be 600 ohms resistance. The line will be distortionless with adjustments calibrated in transmission units and will have a characteristic impedance of 600 ohms resistance.

Such a reference system is now being constructed. The transmitter element consists of a condenser type transmitter and multistage vacuum tube amplifier. The receiver element consists of an amplifier and specially damped receiver. Each element is adjusted to give only negligible distortion over the frequency range.

It is proposed when this system is completed and adjusted that it will be adopted as the Transmission Reference System for telephone transmission work. Other secondary reference systems, employing commercial-type apparatus will be calibrated in terms of the primary system and used for field or laboratory tests when such commercial type systems are needed.

Practical Application of the Recently Adopted Transmission Unit

By C. W. SMITH

THE purpose of this paper is to outline the practical considerations involved in the use of the transmission unit (abbreviated *TU*), which was recently adopted by the Bell System to replace the mile of standard cable in transmission engineering work. A description of the *TU*, together with a discussion of the considerations which led to its adoption has been given by Mr. Martin in another article in this issue.

Effect of Adopting the *TU* as Regards Transmission Standards

The transmission standards in general use vary from 18 miles of standard cable to about 30 miles of standard cable, depending upon the locality and the class of service such as local and toll. It has become customary among telephone people interested in standards of service to associate certain figures for transmission standards with the corresponding standards of service which they represent. It is a distinct advantage, therefore, to retain the same figures for the same standards of service when changing to the new unit. The zero of reference was so selected, therefore, that 21 *TU* is equivalent to 21 miles of standard cable in volume reproduction. This means that if one talks with the same loudness over a circuit of 21 *TU* as over a circuit of 21 miles of standard cable, the volume received from each will be the same. As the attenuation corresponding to the *TU* is only about 6 per cent. less than the attenuation corresponding to the mile of standard cable and 21 miles represents the mean between the highest and lowest standards in common use, transmission standards on the new basis are very little different numerically from the same standards on the old basis. The former 18-mile standard is equivalent in transmission to 17.6 *TU* and the 30-mile standard is equivalent to 30.4 *TU*. The same numerical values can, therefore, generally be used for transmission standards in the new system, as in the old, since the greatest differences encountered will be 0.4 *TU*.

It is also true that a given transmission loss specified in miles will correspond very closely in numerical value to the same loss expressed in *TU*. People not directly engaged in transmission work, therefore, may generally disregard the slight difference which exists in considering transmission losses expressed in *TU* as compared with standard cable.

Use of the *TU* in Transmission Studies

In making transmission studies it has previously been the practice to express the transmission efficiency of limiting subscribers' loops in terms of the resistance of a 22-gauge loop which would have the same total transmitting and receiving loss, thus a 400-ohm loop meant a loop which had the same total transmitting and receiving loss as a loop of 22-gauge ASA cable having a resistance of 400 ohms. At the time of changing from miles to *TU*, it was decided to abandon this method of expressing limiting loop losses in the Bell System and to express them directly in *TU*; thus a 5 *TU* loop means a loop whose total transmitting and receiving loss, taking into account the efficiency of the subscribers' set, is 5 *TU*. The following table gives a number of limiting loops expressed in *TU* and their equivalents in ohms of 22-gauge cable as defined above, assuming the use of the most efficient type of subscriber's set now available.

Limiting Loops Expressed in *TU*	Limiting Loops Expressed in Ohms
3	312
4	350
5	387
6	424
7	461
8	499
9	537
10	573
11	610

Conversion from Miles to *TU* and Computation of Transmission Equivalents

During the transition period in the adoption of the *TU* it will frequently be necessary to convert transmission data which are expressed in miles, to *TU*. This is easily accomplished by multiplying by a conversion factor and in the case of the transmission efficiencies of subscribers' sets by also correcting for the difference in the reference zero which was brought about for reasons referred to above. Two units both known as miles have been in common use as a measure of transmission; they are the standard cable mile and the 800-cycle mile. A different conversion factor is required for each.

The attenuation constant of standard cable for the complex currents used in the transmission of speech varies appreciably with the length of cable considered, since for long lengths the higher frequencies are attenuated to such low values as to have very little effect on the received volume. The best average figure is 0.122, although this value has yet to be determined more precisely by careful laboratory tests. The attenuation corresponding to one *TU* for currents of any

frequency is 0.115. The ratio of the effect on volume of the mile of standard cable to the TU is, therefore, $\frac{0.122}{0.115}$ or 1.06, and equivalents obtained by comparison with standard cable by means of talking tests can therefore be converted to TU by multiplying by this factor, as previously indicated.

The 800-cycle mile which has been commonly used in expressing computed transmission losses, has an attenuation of 0.109 to currents of any frequency, and therefore data expressed in 800-cycle miles are converted to TU by multiplying by $\frac{0.109}{0.115}$ or 0.95.

For making talking tests the field has been supplied with artificial cables which were slightly different from standard cable, having a capacity of .06μf. per mile instead of .054 μf. Miles of this artificial cable may be converted to TU by multiplying by 1.12.

The conversion of subscribers' loop losses to TU is somewhat more complicated as the zero of reference for subscriber's set efficiencies is slightly different on the new basis. In the Bell System, therefore, complete data on subscribers' loop losses in terms of the new unit were made available for engineering work at the time the TU was adopted.

The transmission equivalent of a line per unit of length in TU may be obtained by multiplying the attenuation constant of the line computed in the usual manner by a conversion factor. Calling the computed attenuation constant of the line per unit of length α, the number of TU will be given by the expression: $TU = \frac{\alpha}{0.115} = 8.69\alpha$.

In finding the total loss which a short line or a piece of equipment, such as, for example, a repeating coil will cause when inserted in a given circuit, the current in the receiving apparatus is usually computed for a convenient voltage applied at the sending end of the circuit, first with the repeating coil in the circuit and then with it out, the applied voltage remaining constant. Calling these currents I_1 and I_2 respectively, the current ratio $\frac{I_1}{I_2}$ may be converted into TU by the expression

$$\text{Loss in } TU = 20 \log_{10} \frac{I_1}{I_2}.$$

Transmission Maintenance

The transmission measuring sets used for checking up the maintenance [1] of the plant from a transmission standpoint, have previously

[1] See an article in this issue "Electrical Tests and Their Applications in the Maintenance of Telephone Transmission"—W. H. Harden.

been calibrated in 800-cycle miles, and as the *TU* is of the same nature as this unit, no difficulties are encountered in arranging the sets to read directly in *TU*.

New sets will be manufactured on this basis, but it will, of course, be desirable in order to avoid frequent conversion of data from one unit to the other, to arrange many of the sets which are already in use in the plant to read in *TU*. It is not planned to convert the sets which depend upon ear comparisons, such as the [2] 1-A and 1-B transmission measuring sets and the receiver shunts used in some cases for checking up repeater gains, as the difference when measuring small values is not great and these sets are generally used for a class of work where the required precision is not sufficient to warrant their conversion to the new basis. Visual reading sets, however, such as 2-A, 3-A and 4-A transmission measuring sets and the 2-A repeater gain set, give results which are accurate to about 0.1 *TU* and are usually used for work where a fairly high degree of precision is required. These sets can be changed to read directly in *TU* at a comparatively small expense as it is only necessary to change the calibration of the measuring dials and slide wire potentiometers and the values of certain of the resistances associated with them. The cost of making these changes will be reduced by the fact that it is planned to make certain other desirable changes which will effect improvements in the operation of the sets at the same time. Complete loss data in terms of *TU* which are necessary for checking measured equivalents, have been prepared and will replace the data formerly used.

In toll line maintenance work, record cards are kept which show the layout of toll circuits and the transmission losses of the component parts of each circuit together with the total loss which should be obtained by test if the circuit is not in trouble. In changing over from miles to *TU* these record cards will be revised to show losses in the new unit.

Crosstalk Computations

In handling certain types of crosstalk problems, it has been found convenient to express crosstalk in terms of transmission units rather than crosstalk units. Miles of standard cable have previously been used in such problems. *TU* can be used for this purpose as well as miles and it is somewhat simpler to make the conversion from

[2] See a paper by E. H. Best, "Measuring Methods for Maintaining the Transmission Efficiency of Telephone Circuits," *Journ. A.I.E.E.*, Vol. XLIII, 1924.

crosstalk units to *TU* than from crosstalk units to miles. Crosstalk may be converted from crosstalk units to *TU* as follows:

$$\text{Crosstalk in } TU = 20 \log_{10} \left(\frac{\text{No. of Crosstalk Units}}{10^6} \right).$$

The number of *TU* corresponding to certain numbers of crosstalk units are whole numbers and are therefore, easy to remember as shown in the following table.

Crosstalk in Terms of *TU* Loss	Crosstalk Units
80	100
60	1,000
54 Approx	2,000
40	10,000
20	100,000

Conclusion

From this discussion the conclusion may be drawn that the adoption of the *TU* in place of the mile as the unit of telephone transmission can be readily accomplished in its practical application in the plant. During the transition period, before complete lists of the new data have been compiled, and before the measuring apparatus in use has all been changed to the new basis, frequent conversions between miles and *TU* will be necessary. These conversions can easily be made by multiplying by the proper conversion factor.

Impedance of Loaded Lines, and Design of Simulating and Compensating Networks

By RAY S. HOYT

Synopsis: A knowledge of the impedance characteristics of loaded lines is of considerable importance in telephone engineering, and particularly in the engineering of telephone repeaters. The first half of the present paper deals with the impedance of non-dissipative loaded lines as a function of the frequency and the line constants, by means of description accompanied by equations transformed to the most suitable forms and by graphs of those equations; and it outlines qualitatively the nature of the modifications produced by dissipation. The characteristics are correlated with those of the corresponding smooth line.

The somewhat complicated effects produced by the presence of distributed inductance are investigated rather fully. In the absence of distributed inductance a loaded line would have only one transmitting band, extending from zero frequency to the critical frequency. Actually, however, every line—even a cable—has some distributed inductance; and the effect of distributed inductance, besides altering the nominal impedance and the critical frequency, is to introduce into the attenuating range above the critical frequency a series of relatively narrow transmitting bands—here termed the "minor transmitting bands"—spaced at relatively wide intervals. The paper is concerned primarily with the impedance in the first or major transmitting band; but it investigates the minor transmitting bands sufficiently to determine how they depend on the distributed inductance, and to derive general formulas and graphical methods for finding their locations and widths—an investigation involving rather extensive analysis.

The latter half of the paper describes various networks devised for simulating and for compensating the impedance of loaded lines; it furnishes design-formulas and supplementary design-methods for all of the networks depicted; and outlines a considerable number of applications pertaining to lines and to repeaters.

Introduction

THE present paper on periodically loaded lines (of the series type) is to some extent a sequel to a previous paper on smooth lines.[1]

The reader may be reminded that the transmission of alternating currents over any transmission line between specified terminal impedances depends only on the propagation constant and the characteristic impedance of the line. In this sense, then, the characteristics of transmission lines may be classed broadly as propagation characteristics and impedance characteristics. In telephony we are concerned primarily with the dependence of these characteristics on the frequency, over the telephonic frequency range.

Prior to the application of telephone repeaters to telephone lines the propagation characteristics of such lines were more important than

[1] "Impedance of Smooth Lines, and Design of Simulating Networks," this *Journal*, April, 1923. Two typographical errors in that article may here be noted: p. 37, formula for C_1, C_2, affix an exponent "2" to the last parenthesis; p. 39, value for C_1, replace comma by decimal point.

their impedance characteristics, because the received energy depended much more on the former than on the latter. Indeed, the object of loading [2] was to improve the propagation characteristics of transmission lines; the effects on the impedance characteristics were incidental, and of quite secondary importance.

The application of the two-way telephone repeater greatly altered the relative importance of these two characteristics, decreasing the need for high transmitting efficiency of a line but greatly increasing the dependence of the results on the impedance of the line. As well known, this is because the amplification to which a two-way repeater can be set without singing, or even without serious injury to the intelligibility of the transmission, depends strictly on the degree of impedance-balance between the lines or between the lines and their balancing networks. In the case of the 21-type repeater the two lines must closely balance each other throughout the telephonic frequency range. In the case of the 22-type repeater, which for long lines requiring more than one repeater is superior to the 21-type, impedance-networks are required for closely balancing the impedances of the two lines throughout the telephonic frequency range. Such balancing networks are necessary also in connection with the so-called four-wire repeater circuit.[3]

In Parts I, II, and III of this paper there is presented in a simple yet fairly comprehensive manner the dependence of the characteristic impedance of periodically loaded lines (of the series type) on the frequency and on the line constants, by means of description accompanied by equations transformed to the most suitable forms and by graphs of those equations. Also, the dependence of the attenuation constant on the frequency is presented to the extent necessary for exhibiting the disposition of the transmitting and the attenuating bands and thus enabling the characteristic impedance to be described with reference to those bands, and the important correlation between the characteristic impedance and the attenuation constant thereby exhibited; for the characteristic impedance by itself is not fully significant.

Parts IV to VIII, inclusive, relate to the simulation and the compensation of the impedance of periodically loaded lines by means of

[2] For the fundamental theory of loaded lines, reference may be made to the original papers of Pupin and of Campbell (Pupin: Trans. A. I. E. E., March 22, 1899 and May 19, 1900; *Electrical World*, October 12, 1901 and March 1, 1902. Campbell: *Phil. Mag.*, March, 1903.)

[3] Regarding the broad subject of repeaters and repeater circuits, reference may be made to the paper by Gherardi and Jewett: "Telephone Repeaters," Trans. A. I. E. E., 1919, pp. 1287–1345.

the simulating and the compensating[4] networks for loaded lines devised by the writer at various times within about the last twelve years. Of course, the impedance of any loaded line could be simulated, as closely as desired, by means of an artificial model constructed of many short sections each having lumped constants; but such structures would be very expensive and very cumbersome. Compared with them the networks described in this paper are very simple non-periodic structures that are relatively inexpensive and are quite compact; yet the most precise of them have proved to be adequate for simulating with high precision the characteristic impedance of any periodically loaded line, while even the least precise (which are the simplest) suffice for a good many applications. The compensating networks also are of simple form. Design-formulas are included for all of the networks depicted; and certain supplementary design-methods are indicated. Finally, a considerable number of practical applications are outlined (Part VIII).

PART I

Impedance of Loaded Lines—General Considerations

Before proceeding to the more precise and detailed treatment of the impedance of periodically loaded lines in Parts II and III, it seems desirable to furnish a background by outlining broadly the salient facts. For this purpose the loaded line will be compared with its "corresponding smooth line," that is, the smooth line having the same total constants (inductance, capacity, resistance, leakance).

Comparison with the Corresponding Smooth Line

At sufficiently low frequencies the impedance of a periodically loaded line approximates to that of the corresponding smooth line;[4] but at higher frequencies departs widely. Moreover, the impedance of the loaded line depends very much on its relative termination—fractional end-section or end-load ("load" is here used with the same meaning as "load coil" or "loading coil").

To bring out simply and sharply the contrast between a periodically loaded line and the corresponding smooth line, the effects of dissipation will at first be ignored, although the contrast is somewhat heightened thereby.

It will be recalled that the attenuation constant, the phase velocity, and the characteristic impedance of a non-dissipative smooth line are

[4] Defined in the second paragraph of Part IV.

independent [5] of frequency; such a line having a transmitting band (that is, a non-attenuating band) extending from zero frequency to infinite frequencies, and a characteristic impedance which is a pure and constant resistance.

In contrast, the corresponding characteristics of a non-dissipative periodically loaded line depend very greatly on the frequency; such a line has an infinite sequence of alternate transmitting and attenuating bands* wherein the impedance varies enormously with frequency, while at the transition frequencies its nature undergoes a sudden change. In this connection it may be remarked that, because of its special practical importance in being the upper boundary frequency of the first or principal transmitting band, the lowest transition frequency is termed the "critical frequency" to distinguish it from the other transition frequencies; though in its essential nature each transition frequency is a "critical" frequency. In the ordinary case, where the distributed inductance is small compared with the load inductance, each transmitting band is very narrow compared with the succeeding attenuating band. In the limiting case of no distributed inductance there is only one transmitting band and one attenuating band, the former extending from zero frequency to the critical frequency and the latter from the critical frequency to infinite frequencies.

The characteristic impedance of any non-dissipative transmission line is or is not pure reactance according as the contemplated frequency is in an attenuating band or in a transmitting band. For in an attenuating band the line cannot receive energy, since it cannot dissipate any energy and cannot transmit any energy to an infinite distance; while in a transmitting band the line must receive energy, because it does transmit. Thus, at the transition frequency between an attenuating band and a transmitting band the characteristic impedance undergoes a sudden change in its nature; the frequency-derivative of the impedance (namely, the derivative of the impedance with respect to the frequency) is discontinuous, so that the graph of the impedance has a corner (salient point) at a transition frequency. Moreover, at certain of the transition frequencies of a non-dissipative periodically loaded line the impedance is zero, and at others is infinite. The mid-point impedances are pure resistances throughout every transmitting band. (The "mid-point" terminations are "mid-load" and "mid-section," that is, "half-load" and "half-section" respectively.)

[5] Except for slight change of the inductance, and even of the capacity, with frequency.

* For distinction, the first (lowest) or principal transmitting band may be termed the "major" transmitting band; the others, the "minor" transmitting bands.

Clearly the characteristic impedance of any dissipative line cannot be pure reactance at any frequency; for the line receives at its sending end the energy dissipated within itself. Also, the presence of dissipation renders the frequency-derivative of the impedance continuous at all frequencies; that is, it rounds off the corners on the graph of the impedance. Dissipation prevents the impedance from becoming either zero or infinite at any frequency; and in general it prevents the mid-point impedances from being pure resistances in the transmitting bands.

In the neighborhood of the transition frequencies of the loaded line, the effects of even ordinary amounts of dissipation may be very large, thus preventing the impedance from attaining the very extreme values of the non-dissipative line; but with that exception it may be said that the contrast between a loaded line and the corresponding smooth line is merely softened or dulled by the presence of ordinary amounts of dissipation: The impedance of the smooth line is no longer pure resistance, and it varies somewhat or even considerably with the frequency.[1] The impedance of the loaded line no longer varies quite so rapidly with the frequency nor attains such extreme values; but, except at low frequencies, it continues to depart widely from the impedance of the corresponding smooth line, and to vary much more rapidly than the smooth line with frequency, besides varying greatly with its relative termination (fractional end-section or end-load).

Non-Dissipative Loaded Lines

Except in the neighborhood of zero frequency and of the transition frequencies, the characteristic impedance of an efficient loaded line is dependent mainly on the inductance and capacity, only relatively little on the wire resistance and load resistance, and very much less still on the leakance. The present paper is confined mainly to non-dissipative loaded lines; it deals first with the limiting case of no distributed inductance, and then with the case where distributed inductance is present. By the neglect of all dissipation the number of independent variables is sufficiently reduced to enable a comprehensive, though only approximate, view to be obtained of the characteristic impedance of loaded lines. Such a view is a valuable guide in engineering work even though in most cases it may be necessary, for final calculations or verifications, to resort to exact formulas (Appendix D) or graphs thereof.

Notation and Terminology

The meanings of the fundamental symbols employed in this paper can be readily seen from inspection of Fig. 1. Thus, C and L denote the capacity and the inductance of each whole section between loads, and L' the inductance of each whole load; the ratio $L:L'$ is denoted by λ. Figs. 1a and 1b represent infinitely long loaded lines terminating

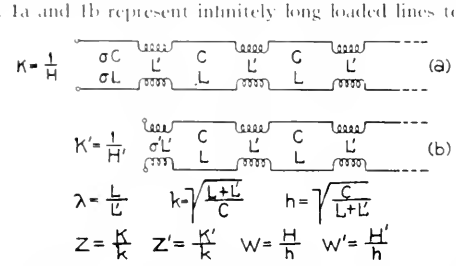


Fig. 1. A Non-Dissipative Infinitely Long Loaded Line Terminating at: (a) σ Section, (b) σ'-Load

at σ-section and σ'-load respectively; the ratios σ and σ' will be termed the "relative terminations." K and K' denote the corresponding characteristic impedances, and H and H' the characteristic admittances. Stated more fully, K denotes the σ-section characteristic impedance, and K' the σ'-load characteristic impedance; similarly for the admittances H and H'. The "nominal impedance" and the "nominal admittance" are denoted by k and h, respectively; that is,

$$k=1\ h=\sqrt{(L+L')\ C}=\sqrt{\ (1+\lambda)L'\ C}, \tag{1}$$

the nominal impedance of a periodically loaded line being defined as equal to the nominal impedance of the corresponding smooth line.[1] $Z=X+iY$ and $Z'=X'+iY'$ denote relative impedances and $W=U+iV$ and $W'=U'+iV'$ the corresponding relative admittances, as defined by the equations

$$Z=K/k, \qquad Z'=K'/k, \qquad W=H/h, \qquad W'=H'/h; \tag{2}$$

the real components being X, X', U, U', and the imaginary components Y, Y', V, V', respectively. By (2),

$$ZW=Z'W'=KH=K'H'=1. \tag{2.1}$$

r denotes the relative frequency, namely, the ratio of any frequency $f=\omega/2\pi$ to the critical frequency f_c; that is, $r=f/f_c=\omega/\omega_c$. i denotes the imaginary operator $\sqrt{\ -1}$.

Besides depending on the frequency f, the quantities K, H, Z, W and K', H', Z', W' depend on the relative terminations σ and σ' respectively (Fig. 1). This dependence will not usually need to be indicated explicitly, but in case of such need the subscript notation will be found convenient. Thus, K_σ will denote the σ-section characteristic impedance (Fig. 1a); and $K_{1-\sigma}$ the "complementary characteristic impedance," that is, the characteristic impedance of the same loaded line if beginning at the "complementary termination"—namely, $(1-\sigma)$-section. As an application of this notation we may note here the relations

$$K_0 = K_1', \qquad H_0 = H_1', \qquad K_1 = K_0', \qquad H_1 = H_0'; \tag{2.2}$$

the first two relations subsisting because of the coincidence of the points σ-section and σ'-load for $\sigma = 0$ and $\sigma' = 1$, and the second two because of the coincidence for $\sigma = 1$ and $\sigma' = 0$.

PART II

Impedance of Non-Dissipative Loaded Lines without Distributed Inductance

Transmitting Band and Attenuating Band

As already stated, a periodically loaded line without distributed inductance (Fig. 1, with $L = 0$) has only one transmitting band and only one attenuating band; the former extending from zero frequency to the critical frequency f_c, and the latter from the critical frequency to infinite frequencies. The formula for f_c is

$$f_c = 1 \div \pi \sqrt{L'C}, \tag{3}$$

L' denoting the inductance of each load and C the capacity of each line-section between loads.

From the energy considerations already adduced, it is known that the characteristic impedance must be pure reactance throughout the attenuating band, but cannot be pure reactance anywhere in the transmitting band.

Formulas for the Relative Impedances

The impedance of even a loaded line without distributed inductance (Fig. 1, with $L = 0$) depends on no less than four independent variables —namely, the frequency f, load inductance L', section-capacity C, and one or the other of the relative terminations σ and σ'. But it is found that these quantities enter in such a way that the relative

impedances $Z = K/k$ and $Z' = K'/k$ and the relative admittances $W = H/h$ and $W' = H'/h$ depend on only two ratios, namely, the relative frequency $r = f/f_0$, and the appropriate relative termination σ or σ', as expressed by the equations[6]

$$Z = \frac{1}{W} = \frac{1}{\sqrt{1 - r^2 + i(2\sigma - 1)r}} = \sqrt{\frac{1 - r^2 + i(1 - 2\sigma)r}{1 - 4\sigma(1 - \sigma)r^2}}, \tag{4}$$

$$Z' = \frac{1}{W'} = \sqrt{1 - r^2 + i(2\sigma' - 1)r} = \sqrt{\frac{1 - 4\sigma'(1 - \sigma')r^2}{1 - r^2 + i(1 - 2\sigma')r}}. \tag{5}$$

In particular, for $\sigma = 0.5$ and $\sigma' = 0.5$, respectively,

$$Z_{.5} = 1/W_{.5} = 1/\sqrt{1 - r^2}, \tag{6}$$

$$Z'_{.5} = 1/W'_{.5} = \sqrt{1 - r^2}. \tag{7}$$

Equations (4) and (5) are not restricted to values of σ and σ' less than unity. On the contrary they are valid for any (real) values of these quantities—though values much exceeding unity are of infrequent occurrence in practice.

Miscellaneous Properties and Relations

Some of the most useful and interesting simple facts deducible from equations (4) and (5) are noted in the next five paragraphs:

In agreement with the general conclusion already reached from energy considerations, equations (4) and (5) show that each of the relative impedances and relative admittances is pure imaginary in the attenuating band ($r > 1$). In the transmitting band ($0 < r < 1$), each is seen to be complex for all values of the relative terminations (σ and σ'), except that each degenerates to a real value when the relative termination becomes 0.5.

Throughout the transmitting band ($0 < r < 1$), a certain conjugate property is possessed by each of the quantities Z, W, Z', W'—namely, each changes merely to its conjugate when σ is changed to $1 - \sigma$, as is readily seen from (4) and (5); that is,

$$Z_\sigma = \bar{Z}_{1-\sigma}, \qquad W_\sigma = \bar{W}_{1-\sigma}, \qquad Z'_\sigma = \bar{Z}'_{1-\sigma}, \qquad W'_\sigma = \bar{W}'_{1-\sigma}, \tag{8}$$

the bar over a symbol denoting the conjugate of the same symbol without the bar. Thus, complementary characteristic impedances are mutually conjugate throughout the transmitting band.

At all values of r,

$$W_\sigma + W_{1-\sigma} = 2W_{.5}, \qquad Z'_\sigma + Z'_{1-\sigma} = 2Z'_{.5}; \tag{9}$$

[6] The equations were written in this sequence because, in practice, section-termination occurs much more frequently than load-termination.

although relations of this form do not hold for Z and for W''. Each of the relations (8) and (9) can be inferred also from simple physical considerations.

Equations (4) and (5) show that W and Z' are alike in form, and also W'' and Z, when σ and σ' are regarded as corresponding to each other; in fact, when $\sigma = \sigma'$,

$$ZZ' = WW'' = W,\ Z' = W'',\ Z = KK',\ k^2 = HH',\ h^2 = 1. \tag{10}$$

Besides, there is the set of perfectly general relations (2.1), which, of course, continue to hold when $\sigma = \sigma'$.

Equations (4) and (5) show also the existence of the following more special relations, holding when the relative terminations (σ and σ') have the values 0 and 1, as indicated by the subscripts:

$$Z_0 Z_1 = Z_0' Z_1' = W_0 W_1 = W_0' W_1' = 1, \tag{11}$$

$$Z_0 = Z_1 = |Z_0'| = |Z_1'| = |W_0| = |W_1| = |W_0'| = |W_1'| = 1. \tag{12}$$

Graphical Representations

Graphical representations of the relative impedances $Z = X + iY$ and $Z' = X' + iY'$, based on equations (4) and (5), will be taken up in the following paragraphs. Evidently it will not be necessary to consider also the relative admittances $W = U + iV$ and $W'' = U' + iV'$ explicitly, since these are of the same functional forms as Z' and Z respectively—as noted in connection with equation (10).

One graphical method of representing the dependence of Z on r and σ is by means of a network of equi-r and equi-σ curves of Z in the Z-plane; likewise the dependence of Z' on r and σ', by means of the equi-r and equi-σ' curves of Z'. The analytic-geometric properties of these curves, as deduced from equations (4) and (5), may be formulated as follows, for any (real) values of σ and σ' but for r restricted to the range 0 to 1:

(*a*) r fixed, σ varied: Z moves on the circle

$$(X - 1)^2 + Y^2 = 1 - (1 - r^2),$$

of radius $1 - \sqrt{1 - r^2}$ with center at $Z = 1 - \sqrt{1 - r^2}$.

(*b*) σ fixed, r varied: Z moves on the curve

$$(X^2 + Y^2)^2 - X^2 - Y^2 (2\sigma - 1)^2 = 0.$$

(*c*) r fixed, σ' varied: Z' moves on the straight line

$$X' = \sqrt{1 - r^2},$$

which is parallel to the Y'-axis at a distance $X' = r^2$ therefrom.

(d) σ' fixed, r varied: Z' moves on the ellipse

$$(X' - 1)^2 + Y'^2/[2\sigma' - 1]^2 = 1,$$

whose center is at $Z' = 0$ and whose semi-axes along the X' and Y' axes have the lengths 1 and $2\sigma' - 1$ respectively.

For values of r, σ, σ' each between 0 and 1, these facts are exhibited graphically in Fig. 2. This is a complex-plane chart of the equi-r

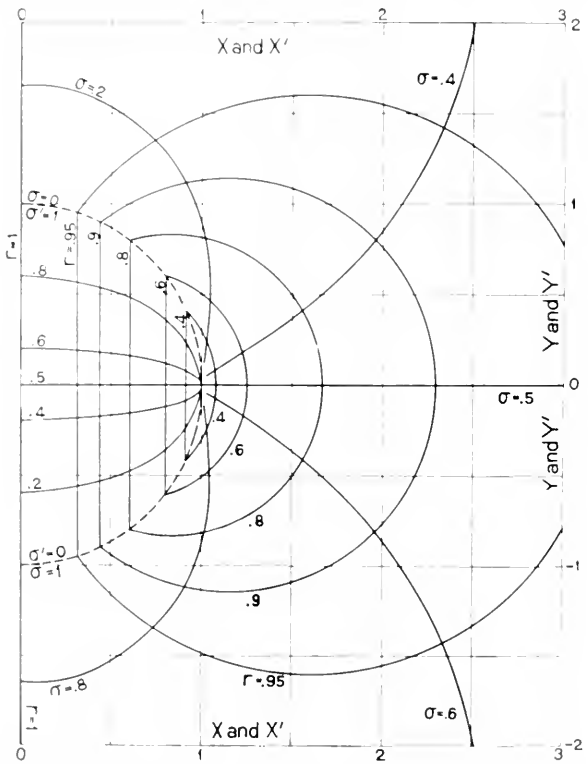


Fig. 2—Complex Plane Chart of the σ-Section Relative Impedance $Z = X + iY$ and the σ'-Load Relative Impedance $Z' = X' + iY'$

and the equi-σ curves of Z, and the equi-r and the equi-σ' curves of Z'. The equi-r and the equi-σ curves constitute a curvilinear network superposed on the rectangular background of $Z=X+iY$; for any assigned pair of values of r and σ the value of Z can be obtained by finding the intersection of those particular curves of r and σ, and at that point reading off the value of Z on the rectangular background. Similarly for the evaluation of Z' by means of the network of equi-r and equi-σ' curves.

For the σ'-range and the σ-range contemplated in Fig. 2—namely, $0<\sigma'<1$ and $0<\sigma<1$—the Z'-realm and the Z-realm are distinct; their mutual boundary (drawn dashed) is the unit semi-circle, that is, the semi-circle of unit radius having its center at the origin. The Z'-realm is the region inside; the Z-realm is all the region outside, extending to infinity in all directions through the positive real half of the complex-plane.

If the ranges of σ' and σ are extended to include values exceeding unity, the Z'-realm and the Z-realm will cease to be distinct but will overlap. The Z'-realm will expand upward, beyond the unit semi-circle, and ultimately will fill the region of unit width extending upward to infinity; the Z-realm will expand into and ultimately will fill the lower half of the unit semi-circle. Hence for values of σ' and σ exceeding unity it is preferable to employ individual charts in representing Z' and Z.

In the language of function-theory it may be said that, when $\sigma'=\sigma$, the Z'-realm and the Z-realm are inverse realms with respect to the unit semi-circle. The straight lines and the circles are inverse curves; the ellipses, and the curves characterized by the equation $(X^2+Y^2)^2-X^2-Y^2/(2\sigma-1)^2=0$ are also inverse curves.

For $r=0$ it is seen that $Z'=Z=1$ for all values of σ' and σ.

For values of r equal to or greater than unity, Z' and Z are pure imaginary, for all values of σ' and σ. For $r=1$, Z' lies somewhere on that part of the imaginary axis constituting the vertical diameter of the unit semi-circle, its position thereon depending on the particular value of σ' contemplated; while Z lies somewhere on the remainder of the imaginary axis. When r approaches infinity, Z' approaches infinity and Z approaches zero, along the imaginary axis.

Another graphical method of representing the relative impedances $Z=X+iY$ and $Z'=X'+iY'$, based on equations (4) and (5), is by means of the Cartesian curves of the components X, Y and X', Y', with the relative frequency r taken as the independent variable and the relative termination (σ or σ') as the parameter.

In this way, Fig. 3 represents X' and Y', and Fig. 4 represents X and Y, all to the same scale. In each of these figures the r-range is 0 to 1.5, thus including the entire transmitting band and a portion of the attenuating band half as wide as the transmitting band. In the

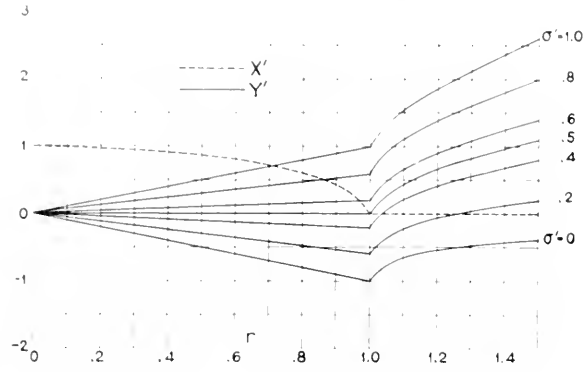


Fig. 3.—Components of the σ-load Relative Impedance $Z' = X' + iY'$.

attenuating band, Z' and Z are pure imaginary; in the transmitting band they are complex in general, though real for $\sigma' = 0.5$ and $\sigma = 0.5$.

Because in practical applications the transmitting band is much more important than the attenuating band, Fig. 5 has been supplied in order to represent X and Y in the transmitting band only, but to a considerably larger scale and for more values of σ.

If σ is read for σ', Fig. 3 will represent U' and V' instead of X' and Y' respectively. If σ' is read for σ, Fig. 4 will represent U' and V' instead of X and Y, so also will Fig. 5.

From Fig. 5 it will be observed that, in a certain range of σ, each curve of X has a maximum at some point within the transmitting band ($0 < r < 1$). For any fixed value of σ (in the range found below) the corresponding maximum of X and the particular value of r (critical value) at which the maximum occurs are expressed by the formulas

$$\text{Max. } X = 1 - 4(1-2\sigma)\sqrt{\sigma(1-\sigma)},$$

$$\text{Crit. } r = \sqrt{\frac{8\sigma(1-\sigma)-1}{4\sigma(1-\sigma)}},$$

as is readily found from the formula for X—namely, the real part of formula (4). The formula for Crit. r shows that the σ-range in which

X, regarded as a function of r, has a maximum within the transmitting band $(0<r<1)$ is

$$(\sqrt{2}-1)\ 2\sqrt{2}<\sigma<(\sqrt{2}+1)\ 2\sqrt{2},$$

that is, approximately,

$$0.116<\sigma<0.854.$$

For values of σ outside of this range, X has no maximum within the transmitting band; but X has then its largest value at $r=0$, decreasing from 1 at $r=0$ to 0 at $r=1$. When $\sigma=1$ 2, Crit. $r=1$;

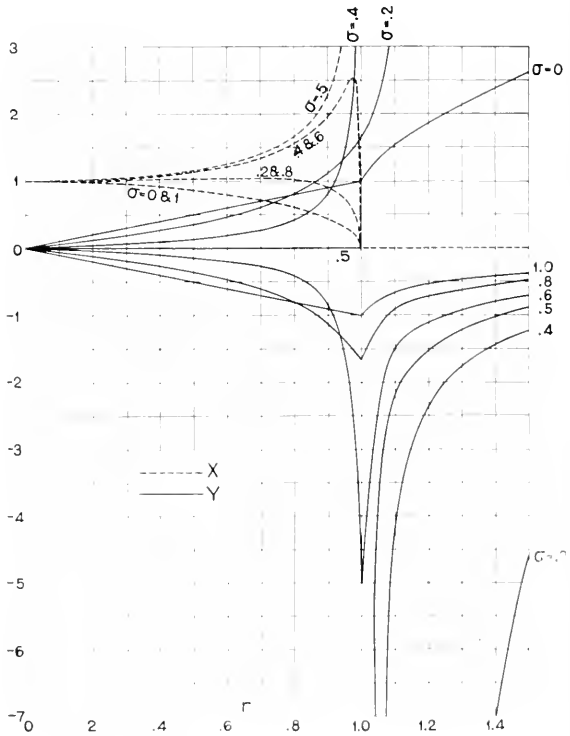


Fig. 4—Components of the σ-Section Relative Impedance $Z=X+iY$

when σ ranges from 1.2 to either of its extreme values appearing in the foregoing inequality for σ, Crit. r decreases from 1 to 0.

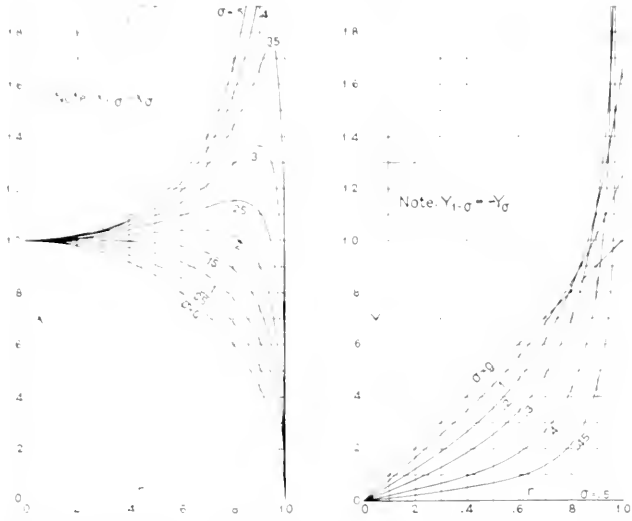


Fig. 5—Components of the σ-Section Relative Impedance $Z = X + iY$ in the Transmitting Band

PART III

Impedance of Non-Dissipative Loaded Lines with Distributed Inductance

Disposition of the Transmitting and the Attenuating Bands

It will be recalled that a loaded line without distributed inductance has only one transmitting band and only one attenuating band. In contrast, a loaded line (Fig. 1) with distributed inductance L has (as shown in Appendix A) an infinite sequence of alternate transmitting and attenuating bands; beginning with a transmitting band extending upward from zero frequency to the first transition frequency which, because of its special practical importance in being the upper boundary frequency of the first or principal transmitting band, is termed the "critical frequency" to distinguish it from the other transition fre-

quencies. The critical frequency will be denoted by f_c; also by f_1—particularly when regarded as the first transition frequency. The relative frequency will be denoted by r, that is,

$$r = f/f_c = f\ f_1. \tag{13}$$

Evidently $r_1 = 1$. General formulas for all of the transition frequencies are furnished a little further on. For the case of no distributed inductance ($L = 0$), there is only one transition frequency—the critical frequency—and it has the value expressed by equation (3). When necessary for distinction, the critical frequency for the case of no distributed inductance will be denoted by f_c', also by f_1'; thus,

$$f_c' = f_1' = 1\ \pi\sqrt{L'C}. \tag{14}$$

The ratio of the critical frequency of any loaded line to the critical frequency of the same loaded line without distributed inductance ($L = 0$) will be denoted by p; that is,

$$p = f_c\ f_c' = f_1\ f_1'. \tag{15}$$

p can be evaluated by means of formula (22).

It is convenient to employ the term "compound band" to denote the band consisting of a transmitting band and the succeeding attenuating band. It is shown in Appendix A that, for any specific loaded line, the widths of all the compound bands are equal; though the transmitting bands become continually narrower with increasing

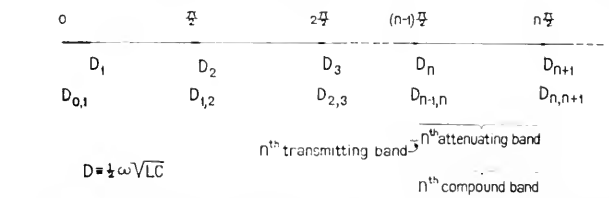


Fig. 6—Scale Showing the Disposition of the Transmitting and the Attenuating Bands of a Periodically Loaded Line (Fig. 1) with Distributed Inductance

frequency, while the attenuating bands become continually wider. These facts are represented on the D-scale in Fig. 6, D being proportional to the frequency f. Fundamentally D denotes the quantity $\frac{1}{2}\omega\sqrt{LC}$; but, by the substitution of $\lambda = L\ L'$, and of r and p defined by (13) and (15), D can be written in the following four identically equivalent forms:

$$D = \tfrac{1}{2}\omega\sqrt{LC} = \tfrac{1}{2}\omega\sqrt{\lambda L'C} = rp\sqrt{\lambda} = rD_1. \tag{16}$$

It is of some interest to note that $D = \frac{1}{2}\omega\sqrt{LC}$ is equal to one-half the "phase constant" ("wave-length constant") of each section of line (L, C) between loads. In Fig. 6 the compound bands are numbered $1, 2, 3, \ldots, n, \ldots$. Thus D_n denotes the transition value of D within the nth compound band; that is, D_n is the value of D at the transition point between the nth transmitting band and the nth attenuating band. $D_{n\,n+1}$ denotes the transition value of D between the nth and $(n+1)$th compound bands; and hence the transition value of D between the nth attenuating band and the $(n+1)$th transmitting band. The corresponding values of f and of ω would be correspondingly subscripted. By (16),

$$D_n = \tfrac{1}{2}\omega_n\sqrt{LC} = \tfrac{1}{2}\omega_n\sqrt{\lambda L'C} = r_n p\sqrt{\lambda} = r_n D_1; \tag{17}$$

and similarly for $D_{n-1,n}$ and $D_{n\,n+1}$. In particular, $D_1 = p\sqrt{\lambda}$, since $r_1 = 1$. As shown in Appendix A,

$$D_{n-1\,n} = (n-1)\pi/2, \qquad D_{n\,n+1} = n\pi/2. \tag{18}$$

Thus the D-width of each compound band is $\pi/2$, that is,

$$D_{n,n+1} - D_{n-1\,n} = \pi/2; \tag{19}$$

and hence, by (16), the f-width has the value

$$f_{n\,n+1} - f_{n-1\,n} = 1/2\sqrt{LC} = 1/2\sqrt{\lambda L'C} = \pi f_1'/2\sqrt{\lambda}. \tag{20}$$

If τ_n denotes the D-width of the nth transmitting band, – that is, $\tau_n = D_n - D_{n-1,n}$, – then the f-width has the value

$$f_n - f_{n-1,n} = \tau_n/\pi\sqrt{LC} = \tau_n/\pi\sqrt{\lambda L'C} = \tau_n f_1'/\sqrt{\lambda}. \tag{20.1}$$

With regard to the nth compound band it will be noted that there are two kinds of transition points – namely, the internal transition point D_n, and the boundary transition points $D_{n-1,n}$ and $D_{n,n+1}$. This distinguishing terminology will be found convenient in connection with the transition frequencies also.

As indicated by Fig. 6, the widths of all the compound bands are equal; but with increasing n the width of the nth transmitting band continually decreases toward a width of 0, while the nth attenuating band continually increases toward a D-width of $\pi/2$; so that the infinitely remote compound bands are pure attenuating bands, the infinitely remote transmitting bands being vanishingly narrow.

The situation of the critical value D_n of D within the nth compound band has no such simple expressions as have the boundary points $D_{n-1\,n}$ and $D_{n\,n+1}$; for D_n is a root of a transcendental equation and can be expressed only by an infinite series of terms or of opera-

tions. In Appendix A a power series formula has been derived for D_n in terms of $\lambda = L\ L'$ and $D_{n-1,n} = (n-1)\pi\ 2$; if, for brevity, the somewhat cumbersome (though expressive) symbol $D_{n\ 1,n}$ is denoted by d_n, this power series is

$$D_n = d_n + \frac{\lambda}{d_n} - \frac{1}{d_n}\left(\frac{\lambda}{d_n}\right)^2 + \left(\frac{2}{d_n^2} - \frac{1}{3}\right)\left(\frac{\lambda}{d_n}\right)^3 - \left(\frac{5}{d_n^3} - \frac{1}{3d_n}\right)\left(\frac{\lambda}{d_n}\right)^4$$
$$+ \left(\frac{14}{d_n^4} - \frac{5}{d_n^2} + \frac{1}{5}\right)\left(\frac{\lambda}{d_n}\right)^5 - \left(\frac{42}{d_n^5} - \frac{56}{3d_n^3} + \frac{23}{15d_n}\right)\left(\frac{\lambda}{d_n}\right)^6 + \ldots, \quad (21)$$

valid for $n = 2,3,4, \ldots$ but not for $n = 1$. For $n = 1$, so that $D_n = D_1$, it is shown in Appendix A that the appropriate formula is [7]

$$D_1 = \sqrt{\lambda}\left(1 - \frac{\lambda}{6} + \frac{11\lambda^2}{360} - \frac{17\lambda^3}{5040} - \frac{281\lambda^4}{604800} + \frac{44029\lambda^5}{119750400} \cdots\right). \quad (22)$$

Since, by (16), $p = D_1 \sqrt{\lambda}$, the series for p is the series in the parenthesis; see also (23-A) in Appendix A. Alternative series-formulas for evaluating D_1 and D_n are derived in Appendix A—formulas (23-A) and (23.1-A) for D_1, and (20.2-A) for D_n. It may be observed that $D_n - d_n < \lambda\ d_n$, that $D_1 < \sqrt{\lambda}$, and that $1 - p < \lambda\ 6$.

The smaller λ, the more convergent are these formulas. Formula (22) is highly convergent, even when λ is as large as unity or even somewhat larger. The convergence of formula (21) depends very much on d_n and hence on n: when n is large, (21) is satisfactorily convergent even for fairly large values of λ; but when n is small, (21) is satisfactorily convergent only for rather small values of λ.

As a supplement to or as an alternative to formulas (21) and (22) there will now be given a widely applicable formula of successive approximation for D_n, valid for all the values of n—including $n = 1$—and suitable even for large values of λ. With $D_n - d_n$ (the D-width of the nth transmitting band) denoted by τ_n, this formula (derived by Newton's general method of approximation) is:

$$\tau_n'' = \frac{\lambda\tau_n' + \lambda \sin \tau_n' \cos \tau_n' - d_n \sin^2 \tau_n'}{\lambda + \sin^2 \tau_n'}, \quad (22.1)$$

wherein τ_n' is some approximate known value of τ_n, and τ_n'' is a more accurate approximate value yielded by the formula. τ_n'', in turn, is to be used in the formula to compute a still more accurate approximate value τ_n'''; and so on, through as many cycles as may be

[7] From the sequence of signs in this formula, namely $-+--+$, the sign of the next term is not evident. A similar remark applies to formulas (23-A) and (23.1-A) in Appendix A.

necessary—usually not more than two or three, though occasionally four. First-approximation values for τ_n are:

$$\tau_n' = \frac{\lambda}{d_n}\left(1 - \frac{\lambda}{d_n'}\right) \text{ when } n = 1,$$

$$\tau_n' = \lambda\,\lambda(1 - \lambda/6) \text{ when } n = 1,$$

as can be seen from (21) and (22) respectively. When $n = 1$, $\tau_n = D_1$, since $d_1 = 0$ by the first of (18).

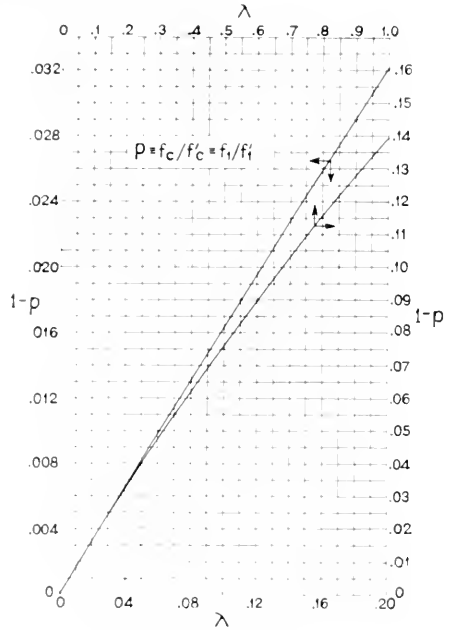


Fig. 7.—Graphs of $1-p$ Representing the Fractional Lowering of the Critical Frequency by Distributed Inductance

D_n having been evaluated, the transition frequency f_n between the nth transmitting band and the nth attenuating band is calculable immediately from

$$f_n = \frac{D_n}{\pi\sqrt{LC}} = \frac{D_n}{\pi\sqrt{\lambda L'C'}} = \frac{D_n f_1'}{\sqrt{\lambda}}, \tag{23}$$

derived from (17) supplemented by (14). Formula (23) is valid also when $n=1$, with D_1 evaluated from one of its appropriate formulas; the resulting formula for the critical frequency $f_1=f_c$ reduces to

$$f_1=f_c=p\lambda\pi\lambda L\bar{C}=p\pi\lambda\sqrt{L'C}=pf_c'=pf_1', \tag{24}$$

because $D_1=p\lambda\lambda$, by (16); it is seen that (24) is consistent with (15).

For use in (24) and for certain other purposes to be met later, Fig. 7 gives graphs of $1-p$, calculated by (22) and also (22.1), for a wide range of λ. Up to the present time the largest value of λ occurring in practical applications in the Bell System is about 0.12; Fig. 7 covers

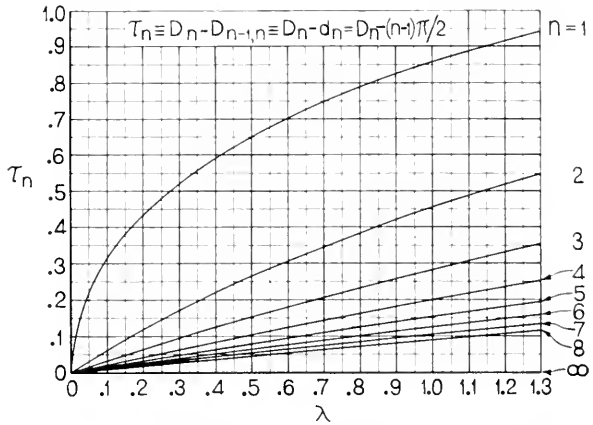


Fig. 7.1 — Graphs for Finding the Widths of the Transmitting Bands

about eight times this range. Inspection of it shows that the graph of $1-p$ is sensibly a straight line up to values of r somewhat larger than even 0.12; and that $1-p$ is only slightly less than λ 6, which is merely the first term in the power series formula for $1-p$ obtained from (22).

The graphs in Fig. 7.1 constructed by means of formulas (22.1), (22), (21) represent directly the dependence of the D-width $\tau_n=D_n-D_{n-1,n}$ of the nth transmitting band on λ and n, for a wide range of λ and the first eight values of n. The f-width is then obtainable

immediately from (20.1); and f_n from (23), since $D_n = \tau_n + (n-1)\pi/2$. In particular, the graph for $n=1$ is a graph of D_1; but D_1 and hence f_1 can be evaluated much more precisely by means of Fig. 7 described in the preceding paragraph.

The boundary transition frequencies $f_{n-1,n}$ and $f_{n,n+1}$ of the nth compound band (any compound band) depend on only one parameter (besides n) namely, the product LC. The internal transition frequency f_n depends on two independent parameters (besides n) namely, the product LC and the ratio $\lambda = L/L'$. Hence, fixing LC fixes all of the boundary frequencies of the compound bands; fixing LC and λ fixes all of the transition frequencies—boundary and internal. Fixing any one boundary frequency fixes LC and thereby fixes all of the remaining boundary frequencies; fixing any two transition frequencies of which at least one is an internal transition frequency fixes LC and λ and thereby fixes all of the remaining transition frequencies—boundary and internal.

The relative widths of all the transmitting and attenuating bands depend on only one parameter—namely, the ratio $\lambda = L/L'$. Hence, fixing λ fixes the relative widths of all these bands; fixing the ratio of the widths of any two bands not both of which are compound bands fixes λ and thereby fixes the relative widths of all the transmitting and attenuating bands.

The effect of increasing λ, when $L'C$ is fixed, is to lower the critical frequency $f_c = f_1$, the critical frequency approaching zero when λ approaches infinity. But for even the largest values of λ met in practice the critical frequency is not much lower than for $\lambda = 0$; the fractional decrease $(f_c' - f_c)/f_c'$ produced in the critical frequency by increasing λ from 0 to any value λ is exactly equal to $1-p$ and hence for any ordinary value of λ is, by (22), closely equal to $\lambda/6$ (which is only 0.02 for $\lambda = 0.12$). It is interesting to note that the nominal impedance—defined by equation (1)—is increased about three times as much as the critical frequency is decreased; for the fractional increase in the nominal impedance is exactly $\lambda/(1+\lambda) = 1$, and hence approximately $\lambda/2$.

All the transition frequencies are reduced by increasing λ, when $L'C$ is fixed. The transition frequencies bounding the compound bands, and hence the widths of the compound bands, decrease in direct proportion to an increase of λ. But the values of the internal transition frequencies do not decrease so rapidly; for the ratio of transmitting band width to attenuating band width increases with increasing λ.

The effect of adding distributed inductance L to a loaded line (L', C) having originally none is to replace the previous single compound band of infinite width by an infinite number of compound bands each of finite width. The larger L the narrower are the compound f-bands, and the further to the left they are situated. Although, as already noted, increasing L decreases the critical frequency, it increases the relative width of each transmitting band—namely, the ratio of the width of each transmitting band to the compound band of which it is a constituent. Thus, when L becomes very large (so that LC and λ become very large) there are within even a moderate frequency-range a very large number of compound bands whose transmitting constituents are very wide compared with the attenuating constituents.

The effect of applying lumped loading to a given smooth line (L, C) is to introduce into the previous transmitting band of infinite width an infinite number of attenuating bands whose upper boundary points are equidistant and whose widths continually decrease toward the lower frequencies. When the inductance L' of the loads is continually increased the attenuating bands continually increase in width as a consequence of their lower boundary points moving downward to lower frequencies, so that ultimately the attenuating bands fill the entire frequency scale from zero to infinity. An alternative but equivalent statement regarding the effect of applying lumped loading is that the previous pure transmitting bands, each of D-width equal to $\pi/2$, become compound bands whose attenuating constituents continually increase in width when L' is increased.

(The four preceding paragraphs are based on the last five paragraphs of Appendix A.)

In Fig. 6 the transmitting bands are represented as being relatively narrow compared with the attenuating bands. In existing loaded lines this is indeed the case, but it is not an inherent relation: for any number of the transmitting bands can be made wider than the associated attenuating bands by so designing the loading (lumped or smooth or both) as to secure a sufficiently large value of the ratio $\lambda = L/L'$. (However, for any fixed loading and hence a fixed value of λ, there is some frequency beyond which the transmitting bands are narrower than the associated attenuating bands.)

There will now be given two examples illustrating the relations represented in Fig. 6, and illustrating also the applications of certain of the foregoing formulas and graphs.

The first example pertains to a heavily loaded open-wire line of No. 12 N. B. S. gauge, having loading coils of inductance $L' = 0.241$

henry at a spacing of $s = 7.88$ miles. The line has a capacity of $.00835 \times 10^{-6}$ farad and an inductance of .00367 henry, each per mile; whence, for each line-segment between loads, $C = .0658 \times 10^{-6}$ farad and $L = .0289$ henry. Therefore $\lambda = 0.12$. With λ known, the internal transition frequencies f_n (with $n = 1, 2, 3, 4, \ldots$) can be readily evaluated from (23) through the values of D_n obtainable from Fig. 7.1. However, when particularly high accuracy is desired for the first transition frequency f_1 the critical frequency this can be attained by resort to formula (22) or to (22.1), or else to Fig. 7; it is thus found that $1 - p = .0196$, whence $p = 0.9804$, and then $f_1 = 2179$ cycles per second, by (24). The f-width of each compound band is 11161, by (20). The following table shows the locations and widths of the first five ($n = 1, 2, 3, 4, 5$) transmitting bands and associated attenuating bands of this loaded line. The numbers in the columns headed $f_{n-1,n}$ and f_n are the transition frequencies constituting, respectively, the lower and upper boundary points of the transmitting bands; and the numbers in the column headed $f_n - f_{n-1,n}$ are therefore the widths of the transmitting bands. The next to the last column shows the relative widths of the transmitting bands, referred to the first or principal transmitting band whose width is $f_1 - 0 = f_1 = 2179$, the critical frequency being 2179. Similarly, the last column shows the relative widths of the attenuating bands.

n	τ_n	$f_{n-1,n}$	f_n	$f_n - f_{n-1,n}$	$(f_n - f_{n-1,n})/f_1$	$(f_{n,n+1} - f_n)/f_1$
1	.3396	0	2,179	2,179	1.000	3.625
2	.0729	11,464	11,996	532	.215	1.110
3	.0377	22,928	23,203	275	.111	1.514
4	.0253	34,392	34,577	185	.074	1.551
5	.0190	45,856	45,995	139	.056	1.569

It will be observed that the transmitting bands decrease rapidly in width at first, then more and more slowly; and that the associated attenuating bands are relatively very wide. For instance, the second transmitting band (0.215) is only about one-fifth the width of the first (1.000), and the second attenuating band (1.110) is more than twenty times the width of the second transmitting band (0.215).

The second example pertains to a hypothetical, though not necessarily impracticable, loaded line. Before loading, the line is the same as in the first example; but it is very lightly loaded namely, with loading coils of inductance $L' = .0578$ henry at a spacing of $s = 15.76$ miles. Hence, $C = 0.1316 \times 10^{-6}$ farad and $L = .0578$ henry. Therefore

$\lambda = 1$. The following table shows the locations and widths of the first eight transmitting bands and attenuating bands. The critical frequency is $f_1 = 3140$, and the f-width of each compound band is 5732.

n	τ_n	$f_{n-1,n}$	f_n	$f_n - f_{n-1,n}$	$(f_n - f_{n-1,n}) \div f_1$	$(f_{n,n+1} - f_n) \div f_1$
1	.8604	0	3,140	3,140	1.000	.826
2	.4579	5,732	7,403	1,671	.532	1.294
3	.2840	11,464	12,500	1,036	.330	1.496
4	.2008	17,196	17,929	733	.234	1.592
5	.1541	22,928	23,490	562	.179	1.647
6	.1247	28,660	29,115	455	.145	1.681
7	.1046	34,392	34,771	382	.122	1.704
8	.0900	40,124	40,452	328	.105	1.721

Comparison of this table with that of the first example brings out the great diversity between the two examples: the minor transmitting bands in the second example are relatively and absolutely much wider and situated at much lower frequencies than in the first example. In the second example the first or principal transmitting band is somewhat wider than the first attenuating band.

A further application of the foregoing formulas and graphs is to obtain a precise and explicit solution of the important practical problem of loading a given smooth line with lumped loading to secure specified values of the critical frequency f_1 and nominal impedance k. The design-problem consists in determining the requisite values of the load inductance L' and load spacing s in terms of f_1 and k and the known values of the inductance and capacity, L'' and C'', per unit length of the given smooth line. Since $L = sL''$ and $C = sC''$, the solution can be obtained as follows: Substituting $L' = sL''\lambda$ into (1) and solving for λ gives

$$\lambda = \frac{L''C''}{k^2 - L''/C''}.$$

Then D_1 becomes known by means of Fig. 7 or Fig. 7.1 or formula (22) or (22.1). Next, s becomes known from (23) or (24):

$$s = D_1 \pi f_1 \sqrt{L''C''}.$$

Finally, from these formulas for λ and s together with the relation $L' = sL''\lambda$, it follows that

$$L' = \frac{D_1(k^2 - L''/C'')}{\pi f_1 \sqrt{L''/C''}}.$$

The Relative Impedances

The formulas for the impedances and admittances of a non-dissipative periodically loaded line (Fig. 1) with any amount of distributed inductance L will next be set down, and discussed somewhat, with particular regard to the transmitting and the attenuating bands of the loaded line.

As before, it is convenient to deal with the relative impedances Z, Z' and the relative admittances W', W'' defined by equations (2). Special attention is given to the particular values $Z_{\frac{1}{2}}, Z'_{\frac{1}{2}}, W'_{\frac{1}{2}}, W''_{\frac{1}{2}}$ corresponding to mid-point terminations.

It is found that Z, Z', W', W'' can be expressed in terms of three independent quantities—namely, the relative frequency $r = f/f_c$, the inductance ratio $\lambda = L/L'$, and the relative termination σ or σ'. For most applications the quantity $r = f/f_c$ is more significant than any other quantity proportional to the frequency f, and on that score it would be desirable to employ it explicitly in the formulas for the impedances and admittances. However, the formulas are rendered considerably more compact by employing the quantity D defined by equation (16). Whenever desired, D can be expressed in terms of r, λ, and p by means of (16); and thence in terms of r and λ by means of (22).

Because of their special importance the formulas for the mid-point relative impedances and relative admittances will be set down first. From Appendix D these formulas are found to be

$$Z_{\frac{1}{2}} = \frac{1}{W'_{\frac{1}{2}}} = \sqrt{\frac{\lambda}{\lambda+1}}\sqrt{\frac{\lambda+D\cot D}{\lambda-D\tan D}} \tag{25}$$

$$Z'_{\frac{1}{2}} = \frac{1}{W''_{\frac{1}{2}}} = \sqrt{\frac{(\lambda+D\cot D)(\lambda-D\tan D)}{\lambda(\lambda+1)}}, \tag{26}$$

$$= \sqrt{\frac{\lambda^2+2\lambda D\cot 2D-D^2}{\lambda(\lambda+1)}}. \tag{26.1}$$

From these formulas it can be verified that $Z_{\frac{1}{2}}$ and $Z'_{\frac{1}{2}}$ are pure imaginary throughout every attenuating band, and it can be seen that they are pure real throughout every transmitting band.

A study of equations (25) and (26) brings out also the following facts regarding the variation of $Z_{\frac{1}{2}}$ and $Z'_{\frac{1}{2}}$ in the transmitting and the attenuating bands, with increasing frequency:

In the first transmitting band, $Z_{\frac{1}{2}}$ ranges from 1 to ∞, but in all of the other odd transmitting bands it ranges from ∞ to ∞, through finite intervening values; in the even transmitting bands it ranges

from 0 to 0, through finite intervening values. In the odd attenuating bands it ranges from $-i\infty$ to $-i0$; and in the even attenuating bands it ranges from $+i0$ to $+i\infty$.

In the first transmitting band, $Z'_{.5}$ ranges from 1 to 0, but in all of the other transmitting bands it ranges from ∞ to 0. In all of the attenuating bands it ranges from $+i0$ to $+i\infty$.

These facts are illustrated by Fig. 8, which gives graphs of $Z_{.5}$ and $Z'_{.5}$ over a range of three compound bands, as functions of $r=f/f_1=D/D_1$, with $\lambda=0.12$; also with $\lambda=0$, for comparison. On the scale there used, the curves for the two values of λ are indistinguishable

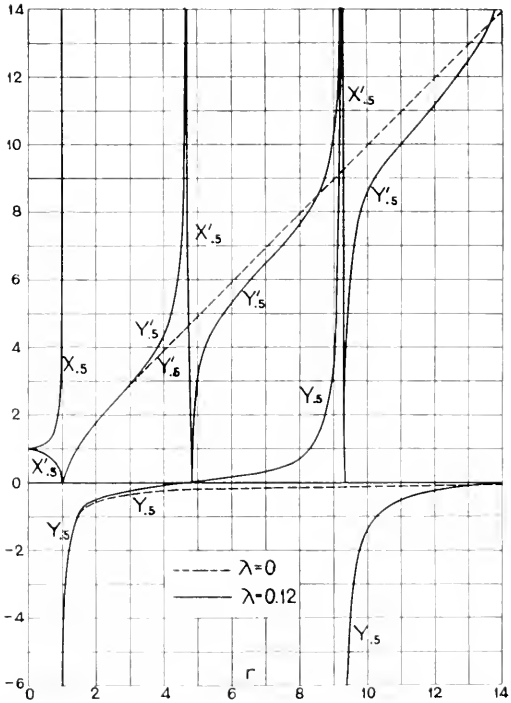


Fig. 8—Mid-Section Relative Impedance $Z_{.5}=X_{.5}+iY_{.5}$ and Mid-Load Relative Impedance $Z'_{.5}=X'_{.5}+iY'_{.5}$ Over a Range of Three Compound Bands

throughout the first transmitting band ($0 < x < 1$) and a considerable part of the succeeding attenuating band; but depart widely beyond.

The exact formulas for Z, W and Z', W' for any terminations σ and σ' can be written in the forms

$$Z = \frac{1}{W} = \frac{Z_s \cot (2\sigma - 1)D + i\sqrt{\lambda(1+\lambda)}}{\cot (2\sigma - 1)D + iZ_s\sqrt{(1+\lambda)\lambda}}, \tag{27}$$

$$Z' = \frac{1}{W'} = Z'_s + i\frac{(2\sigma' - 1)D}{\sqrt{\lambda(1+\lambda)}}. \tag{28}$$

These equations are not restricted to values of σ and σ' less than unity; they are valid for any (real) values of these quantities. When $\lambda = 0$, they reduce immediately to (4) and (5) respectively.

From (27) and (28) it is readily verified that Z and Z' are pure imaginary throughout every attenuating band, and it can be easily seen that they are complex throughout every transmitting band; because Z_s and Z'_s are pure imaginary throughout every attenuating band, and pure real throughout every transmitting band.

It is seen from (27) and (28) that, throughout every transmitting band, each of the quantities Z, W, Z', W' changes merely to its conjugate when σ is changed to $1-\sigma$. Thus the conjugate property expressed by equations (8) is not limited to loaded lines without distributed inductance but holds when there is any amount of distributed inductance. Thus it continues to be true that complementary characteristic impedances are mutually conjugate—throughout every transmitting band. For Z' and W', these facts are readily seen from physcial considerations also; though not so readily for Z and W.

From physical considerations, as well as from equation (28), it is readily seen that Z' continues to possess the property expressed by the second of equations (9); on the other hand, W no longer possesses the property expressed by the first of (9).

We shall now return to the important formulas (25) and (26) for the mid-point relative impedances in order to discuss them for small values of λ such as occur in practice, and particularly for a frequency-range not greatly exceeding that of the first transmitting band. For this purpose it is advantageous to write these formulas in the following forms, notwithstanding some sacrifice of compactness:

$$Z_s = \frac{1}{W_s} = \sqrt{\frac{\lambda + D\cot D}{\lambda + 1}}\left(\sqrt{1 - \frac{D\tan D}{D_1 \tan D_1}}\right), \tag{29}$$

$$Z'_s = \frac{1}{W'_s} = \sqrt{\frac{\lambda + D\cot D}{\lambda + 1}}\left(\sqrt{1 - \frac{D\tan D}{D_1 \tan D_1}}\right). \tag{30}$$

For the discussion of these it should be recalled that $D = \frac{1}{2}\omega\sqrt{\lambda L'C}$ and $r = D/D_1 = f/f_1 = f/f_c$; also that $D_1 \tan D_1 = \lambda$, whence D_1 is approximately equal to $\sqrt{\lambda}$ when λ is small.

Equations (29) and (30) are in such form as to exhibit the manner in which $Z_{.5}$ and $Z'_{.5}$ approach their simple limiting values for $\lambda = 0$, represented by equations (6) and (7) respectively. For when λ approaches 0, $D \cot D$ and $D \tan D$ approach 1 and D^2 respectively; and for values of λ even larger than the largest (about 0.12) occurring in practice, $D \cot D$ and $D \tan D$ respectively are at least roughly equal to 1 and to D^2 throughout even more than the first transmitting band.

The expression for $Z_{.5}$ reduces immediately to $1 - \sqrt{1-r^2}$ when λ is zero. When λ is not zero, $Z_{.5}$ is less than $1 - \sqrt{1-r^2}$ for all values of r in the first transmitting band ($0 < r < 1$); when r increases from 0 to 1, $Z_{.5}$ increases from 1 to ∞.

The expression for $Z'_{.5}$ reduces immediately to $\sqrt{1-r^2}$ when λ is zero. Even when λ is several tenths, $Z'_{.5}$ is very closely equal to $\sqrt{1-r^2}$ for all values of r in the first transmitting band; when r increases from 0 to 1, $Z'_{.5}$ decreases from 1 to 0.

Effects of Distributed Inductance; the "Simulative Loaded Line"

The above-described relations are exemplified in Fig. 9, which gives graphs of $Z_{.5}$ and $Z'_{.5}$ over the first transmitting band and part of the succeeding attenuating band, as functions of r, with λ as parameter equal to 0.12 and to 0. It is seen that the curves of $Z_{.5}$ for the two values of λ do not differ much in the transmitting band ($0 < r < 1$); and that the curves of $Z'_{.5}$ for the two values of λ are indistinguishable—on the scale there used.

In order to indicate more precisely to what extent the forms of $Z_{.5}$ and $Z'_{.5}$ are affected by the presence of distributed inductance, as specified by $\lambda = L/L'$, Fig. 10 has been prepared. This gives a graph of the ratio of the values of $Z_{.5}$ for $\lambda = 0.12$ and $\lambda = 0$; and likewise of $Z'_{.5}$. That is, formulated in functional notation, it gives graphs of $Z_{.5}(r, \lambda)/Z_{.5}(r, 0)$ and $Z'_{.5}(r, \lambda)/Z'_{.5}(r, 0)$. From these it is seen that, in the transmitting band, the mid-section ratio (first ratio) and the mid-load ratio (second ratio) do not differ from unity by more than four per cent. and one-tenth of one per cent., respectively. These observations—particularly the second—suggest that, at least over the whole of the first transmitting band, the impedance of a non-dissipative periodically loaded line with small distributed inductance

can be rather closely simulated by a periodically loaded line without distributed inductance but with suitably chosen load-inductance L_0' and section-capacity C_0. The utility of this observation resides

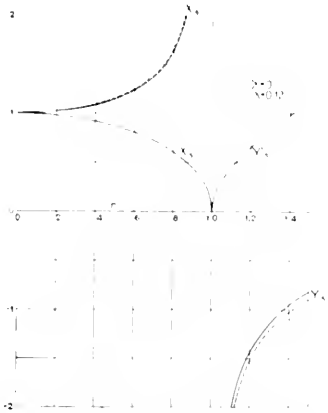


Fig. 9—Mid-Section and Mid-Load Relative Impedances Z_s and Z'_s Over the First Transmitting Band and Part of the Succeeding Attenuating Band

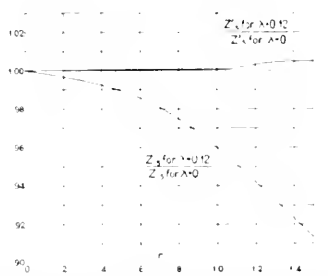


Fig. 10—Ratio Curves Showing Effects of Distributed Inductance on the Forms of the Curves of Z_s and Z'_s.

ultimately in the fact that the formulas for loaded lines without distributed inductance are much simpler than those for loaded lines with distributed inductance.

For mid-section or for mid-load termination the simulation of the effects of distributed inductance described in the preceding paragraph can be made exact at two different frequencies simultaneously, and the requisite values of the load-inductance L_0' and section-capacity C_0 of the simulating loaded line thereby determined. This simulating loaded line will be termed the "simulative loaded line" corresponding to the two particular frequencies contemplated.

In many applications a suitable simulation can be attained by imposing the conditions that the simulating loaded line (L_0', C_0) shall have the same nominal impedance k and critical frequency f_c as the actual loaded line (L', L, C). The particular simulating loaded line so determined will be called the "principal simulative loaded line"; evidently its load-inductance L_0' and section-capacity C_0 are determined in terms of k and f_c and also in terms of L', L, C by the pair of equations

$$k = \sqrt{(L'+L)/C} = \sqrt{L_0'/C_0}, \tag{31}$$

$$f_c = p/\pi\sqrt{L'C} = 1/\pi\sqrt{L_0'C_0}, \tag{32}$$

of which (31) corresponds to (1), and (32) to (15) and (14) combined or to (24). The solution of the pair of equations (31) and (32) is the pair of values

$$L_0' = L'(\sqrt{1+\lambda})/p = k/\pi f_c, \tag{33}$$

$$C_0 = C/p\sqrt{1+\lambda} = 1/\pi f_c k. \tag{34}$$

In conjunction with (22), these formulas show that $L_0' > L'$ and $C_0 < C$; in fact they show that $L_0'/L' = 1+2\lambda/3$ and $C_0/C = 1-\lambda/3$, as first approximations; precise values of these ratios can be readily calculated by substituting for p the power series contained in equation (22).

The simulative precision of the "principal simulative loaded line" depends on the value of the relative termination (σ or σ'). The simulation is far more precise for mid-load termination ($\sigma' = 0.5$) than for mid-section termination ($\sigma = 0.5$); this can be seen by developing in power series the functions involved; for $\lambda = 0.12$ the fact is illustrated by Fig. 10 already cited. The simulative precision for other terminations will not be discussed here, beyond remarking that the "principal simulative loaded line" terminating at σ'-load could not exactly simulate the actual loaded line terminating at σ'-load, even if the simulation were exact at 0.5-load; for the excess-inductances $(\sigma'-0.5)L_0'$ and $(\sigma'-0.5)L'$ are not exactly equal, the former being slightly the larger—as shown by equation (33). However, the smallness of the impedance-departure between the "principal simulative

loaded line" and the actual loaded line when both lines terminate at mid-load can be identically preserved for any other load-point termination of either line by so choosing the load-point termination of the other line that the excess inductance of its end-load beyond half load has the same value. This fact should be kept in mind when designing simulating and compensating networks, particularly such as pertain to a loaded line that terminates with a fractional load; also when choosing the relative termination σ' of the fractional load.

Some idea as to the simulative precision of the propagation constant $\Gamma = A + iB$ of the "principal simulative loaded line" can be obtained from Fig. 22 in Appendix A. For the present purpose the graphs for $\lambda = 0$ can be regarded as pertaining exactly to the "principal simulative loaded line" corresponding to any non-dissipative periodically loaded line having any amount of distributed inductance, while the graphs for $\lambda = 0.12$ are for any non-dissipative loaded line having the particular inductance ratio $\lambda = 0.12$. Of course, A is zero in the range $0 < r < 1$.

PART IV

Networks for Simulating and for Compensating the Impedance of Loaded Lines—General Considerations

The remainder of the paper relates to the simulation and the compensation of the impedance of periodically loaded lines by means of the simulating and the compensating networks devised by the writer, as mentioned in the latter part of the Introduction.

The term "compensating network" requires at least a tentative definition. The compensating networks dealt with in the present paper are of two types: reactance-compensators, and susceptance-compensators. For the present they may be defined—rather narrowly—with reference to the first transmitting band of non-dissipative loaded lines, as follows: a reactance-compensator is a network that neutralizes the characteristic reactance of the line and hence simulates its complementary characteristic reactance; a susceptance-compensator is a network that neutralizes the characteristic susceptance of the line and hence simulates its complementary characteristic susceptance.

As actually worded, this division (Part IV) of the paper pertains mainly to the simulation of loaded lines; but with appropriate slight changes of wording most of it pertains also to compensation. Compensation is dealt with explicitly in portions of Parts V and VIII of the paper.

The simulating and the compensating networks were devised from purely theoretical studies of the characteristic impedance and admittance of periodically loaded lines as dependent on the frequency and on the relative termination, in somewhat the same way as the previously described [1] networks for smooth lines were devised from purely theoretical studies of the characteristic impedance of smooth lines as dependent on the frequency.

Building-out Structures, Basic Networks, and Excess-Simulators

Although the characteristic impedance of a periodically loaded line depends greatly on its relative termination (σ or σ'), yet there is no need of attempting to devise various independent networks corresponding to various relative terminations of the line. For any network that will simulate the line-impedance at any particular relative termination can be "extended" or "built-out" to simulate it at any other relative termination by merely supplementing the network with an "extension network" or "building-out structure" in the nature of an artificial line structure corresponding as closely as may be necessary to the portion of actual line structure included between the two relative terminations contemplated. Simulation can be attained also by building-out the line instead of the network, or by building-out both the line and the network to any common relative termination; but in practice these alternatives are not usually permissible, the usual requirement being the simulation of a given fixed line. (In present practice, the line is terminated usually at midsection [$\sigma = 0.5$], or as closely thereto as practicable.)

The term "basic network" will be used to denote a network which simulates the characteristic impedance of a non-dissipative periodically loaded line without the network's containing in its structure any building-out elements. Regarding the loaded line, the particular relative termination to which the basic network pertains will be termed the "basic relative termination" of the loaded line, and will be denoted by σ_b or σ_b' whenever a symbol is needed for it. (For the kinds of basic networks thus far devised, σ_b and σ_b' lie between about 0.1 and about 0.2, that range having been found to include the relative terminations most favorable to the design of those kinds of basic networks.) The foregoing terms, when used in connection with a dissipative loaded line, will be understood to refer to the corresponding non-dissipative loaded line. A considerable number of kinds of basic networks will be described in Part V supplemented by Part VI.

The amount by which the characteristic impedance of any periodically loaded line exceeds the impedance of the corresponding non-dissipative loaded line will be termed the "excess impedance" (or, more fully, the "excess characteristic impedance"); and a network for simulating it will be termed an "excess-simulator." Excess-simulators for loaded lines will be considered very briefly in Part VII.

(In passing, it may be noted that the foregoing definition of the "excess impedance" of a periodically loaded line properly includes the definition already given [1] of the excess impedance of a smooth

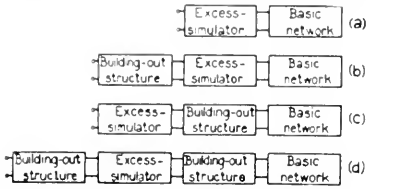


Fig. 11. Abstract Diagrams of Complete Networks for Simulating Characteristic Impedance of Loaded Line

line; for the "nominal impedance" of any smooth line was defined [1] as the impedance of the corresponding non-dissipative smooth line. A similar statement is applicable to the terms "excess simulator" and "basic network" previously defined [1] for smooth lines.)

The foregoing considerations and definitions have prepared the way for Fig. 11, which indicates in an abstract manner how the impedance of any loaded line having any relative termination can be simulated by combinations of basic networks, excess simulators, and building-out structures.

Fig. 11a corresponds to the simple but unusual case in which the loaded line has the basic relative termination: its impedance then can be simulated by the corresponding basic network and excess simulator, without any building-out structure.

When, as usual, the given line does not have the basic relative termination, there are available the two natural alternatives represented by Figs. 11b and 11c. Fig. 11b shows the whole network of Fig. 11a built-out to the relative termination of the given line by means of the requisite building-out structure, which for the highest precision must be dissipative to correspond to the actual line. In Fig. 11c the basic network is built-out to the relative termination of the given line with a non-dissipative building-out structure; and then the resulting network, which simulates the impedance that the actual

line would have if non-dissipative, is supplemented with an excess-simulator such as to simulate the excess impedance of the actual line.

Since the excess impedance depends somewhat on the relative termination it can be simulated more easily at certain relative terminations than at others. This fact is utilized in the arrangement represented by Fig. 11d. Here the basic network is built-out to some relative termination that is particularly favorable for the design of an excess-simulator; the excess-simulator is applied; and then is applied the building-out structure, which for the highest precision must be dissipative to correspond to the actual line.

The simulation-range of the basic networks described in this paper is a little less than the first transmitting band of the loaded line; but after a basic network has been built-out, its simulation-range may extend a little way into the succeeding attenuating band, omitting the immediate neighborhood of the critical frequency. The compensation-range of the compensating-networks is somewhat less than the first transmitting band of the loaded line.

PART V

Networks for Non-Dissipative Loaded Lines without Distributed Inductance

In this Part will be described a considerable number of kinds of "basic networks" for simulating the characteristic impedance of non-dissipative loaded lines without distributed inductance; and two types of compensating networks for such lines. The modifications necessary when the lines have small distributed inductance will be indicated in Part VI.

The various kinds of basic networks here described may be regarded as of two different types corresponding to the terminations of the loaded lines to which they pertain; there may be several varieties of each type. The two types correspond to fractional-section and to fractional-load terminations respectively; that is, to the relative terminations σ_b and σ_b' respectively. (It has been stated already, in Part IV, that σ_b and σ_b' lie between about 0.1 and about 0.2.) It will appear below that these two types are inverse types, in the sense that the impedance of a network of one type is of the same functional form as the admittance of the corresponding network of the other type, when the frequency is regarded as the independent variable. In particular, for equal relative terminations ($\sigma_b = \sigma_b'$), the ratio of the impedance and the admittance of any two corresponding inverse networks is independent of frequency. This corresponds to the relations $Z_1 W'' = 1$ and $Z' W = 1$, holding for the loaded line

itself, according to equations (1) and (5). Hence the two types of networks will sometimes be distinguished as impedance type and admittance type. More specifically, the simulating networks of the two types will be distinguished as impedance-simulators and admittance-simulators, respectively; and the compensating networks as reactance-compensators and susceptance-compensators, respectively.

By being built out to the requisite extent, either type of network evidently can be employed with a loaded line terminating at any point in either a section or a load; but, depending on such termination, one type will require less building-out than the other, and hence will be somewhat preferable on that score. For instance, for simulating the impedance of a loaded line terminating at mid-section ($\sigma = 0.5$), a basic network of the fractional-section type of termination will require less building-out than one of the fractional-load type of termination.

The Basic Networks

The various basic networks mentioned will now be described briefly, by aid of circuit diagrams which show the forms of the networks and which include explicit design-formulas for the proportioning. Mutually corresponding networks of inverse types will be described together or in sequence, in order to exhibit clearly their correlation.

In the design-formulas the requisite values for the network-elements will be expressed in terms of the load-inductance L' and the section-capacity C of the given loaded line; but when desired they can instead be readily expressed in terms of the nominal impedance k and critical frequency f_c, by means of the relations

$$L' = k \ \pi f_c, \qquad C = 1 \ \pi k f_c.$$

Of course, the design-formulas involve also the relative terminations σ and σ'.

Figs. 12 and 13 show two rather simple networks which simulate very well, over most of the transmitting band, the σ-section character-

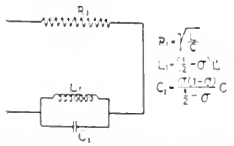


Fig. 12 — Impedance-Simulator for a Loaded Line Terminating at σ-Section, with σ in the Neighborhood of 0.2

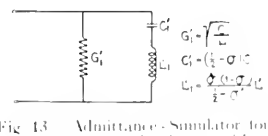


Fig. 13 — Admittance-Simulator for a Loaded Line Terminating at σ'-Load, with σ' in the Neighborhood of 0.2

istic impedance and the σ'-load characteristic admittance, respectively, of a non-dissipative loaded line, when σ and σ' are in the neighborhood of 0.2. The theoretical bases of these two networks and of their proportioning are outlined in Appendix B. (See also Patent No. 1121901 and No. 1437422, respectively.)

Figs. 14 and 15 show two networks which are considerably less simple than those of Figs. 12 and 13 but possess a substantially wider

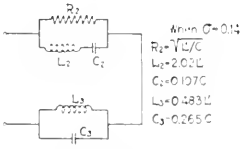


Fig. 14 — Impedance-Simulator for a Loaded Line Terminating at σ-Section, with σ about 0.14

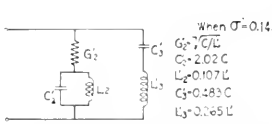


Fig. 15 — Admittance-Simulator for a Loaded Line Terminating at σ'-Load, with σ' about 0.14.

frequency-range of simulation; for them the best value of σ and of σ' is about 0.14. The theoretical bases of these two networks are indicated below in the descriptions of the networks in Figs. 20 and 21, respectively. (See also Patent No. 1167093 and No. 1437422, respectively.)

Fig. 16 shows a network called a reactance-compensator, for a non-dissipative loaded line terminating at σ-section. When proportioned

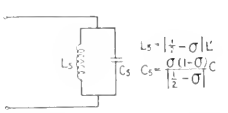


Fig. 16 — Reactance-Compensator for a Loaded Line Terminating at σ-Section: Reactance-Simulator when $0<\sigma<1.2$ Reactance-Neutralizer when $1.2<\sigma<1$

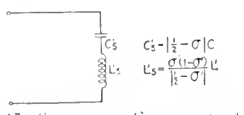


Fig. 17 — Susceptance-Compensator for a Loaded Line Terminating at σ'-Load: Susceptance-Simulator when $0<\sigma'<1.2$ Susceptance-Neutralizer when $1.2<\sigma'<1$

in accordance with the design-formulas there given, this network possesses the following two-fold property with reference to the σ-section characteristic reactance of the loaded line: When σ has any fixed value between 0 and 1.2, the network exactly simulates the σ-section reactance, and exactly neutralizes the $(1-\sigma)$-section reactance; or, what is equivalent, when σ has any fixed value between 1.2 and 1, the network exactly neutralizes the σ-section reactance and exactly simulates the $(1-\sigma)$-section reactance.

Fig. 17 shows a network called a susceptance-compensator, for a non-dissipative loaded line terminating at σ'-load. When proportioned in accordance with the design-formulas there given, this network possesses the following two-fold property with reference to the σ'-load characteristic susceptance of the loaded line: When σ' has any fixed value between 0 and 1.2, the network exactly simulates the σ'-load susceptance, and exactly neutralizes the $(1-\sigma')$-load susceptance; or, what is equivalent, when σ' has any fixed value between 1.2 and 1, the network exactly neutralizes the σ'-load susceptance and exactly simulates the $(1-\sigma')$-load susceptance.

It may be noted that the resonant frequency f_r of the compensators in Figs. 16 and 17 is never less than the resonant frequency f_c of the loaded line; for when $\sigma=\sigma'$ the two types of compensators have the same value of f_r, and

$$f_r \div f_c = 1.2\sqrt{\sigma(1-\sigma)}.$$

This ratio has a minimum value of unity, when $\sigma=1.2$; and becomes infinite when $\sigma=0$ and when $\sigma=1$. It is equal to 1.25 when $\sigma=0.2$ and when $\sigma=0.8$.

The compensators in Figs. 16 and 17 are evidently inverse networks; the theoretical principles underlying them are outlined together in Appendix C. (See also Patent No. 1243066 and No. 1475997, respectively.)

With σ and σ' each in the neighborhood of 0.2 or of 0.8, the σ-section characteristic reactance and the σ'-load characteristic conductance of a non-dissipative loaded line are simulated pretty well by the constant resistance R_1 and the constant conductance G_1' of Figs. 12 and 13, respectively, as pointed out in Appendix B.

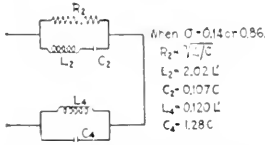


Fig. 18.—Resistance-Simulator for a Loaded Line Terminating at σ-Section, with σ about 0.14 or about 0.86.

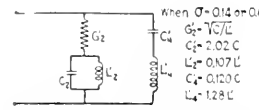


Fig. 19.—Conductance-Simulator for a Loaded Line Terminating at σ'-load, with σ' about 0.14 or about 0.86.

Simulation of the σ-section resistance and of the σ'-load conductance can be accomplished over a substantially wider frequency-range than in the foregoing paragraph, by means of the networks of Figs. 18 and 19, respectively; for them the best value of σ and of σ' is about 0.14.

These networks must not be confused with those of Figs. 14 and 15: they are like the latter in form but differ in the values of certain of their elements, as will be seen on close examination; they differ also in their functions, the networks of Figs. 14 and 15 simulating the σ-section impedance and the σ'-load admittance, respectively, whereas the networks of Figs. 18 and 19 simulate merely the resistance and the conductance components of these, respectively. In Fig. 18 the reactance of the L_4C_4-portion neutralizes that of the $R_2L_2C_2$-portion; and in Fig. 19 the susceptance of the $L_4'C_4'$-portion neutralizes that of the $G_2'C_2'L_2'$-portion. (See also Patent No. 1167693 and No. 1437422, respectively.)

By combining the resistance-simulator of Fig. 18 and the reactance-simulator of Fig. 16 there results the impedance-simulator of Fig. 20.

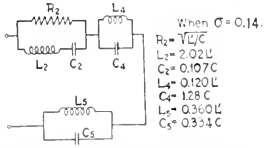


Fig. 20—Impedance-Simulator for a Loaded Line Terminating at σ-Section, with σ about 0.14. (This figure indicates the synthesis of the network in Fig. 14.)

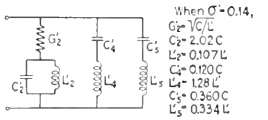


Fig. 21—Admittance-Simulator for a Loaded Line Terminating at σ'-Load, with σ' about 0.14. (This figure indicates the synthesis of the network in Fig. 15.)

But it is found that the L_4C_4-portion and the L_5C_5-portion can be combined, without appreciable sacrifice of simulative precision, into the single L_3C_3-portion of Fig. 14—whose synthesis is thereby indicated. (See also Patent No. 1167693.)

By combining the conductance-simulator of Fig. 19 and the susceptance-simulator of Fig. 17 there results the admittance-simulator of Fig. 21. But it is found that the $L_4'C_4'$-portion and the $L_5'C_5'$-portion can be combined, without appreciable sacrifice of simulative precision, into the single $L'_3C''_3$-portion of Fig. 15 whose synthesis is thereby indicated. (See also Patent No. 1437122.)

PART VI

Networks for Non-Dissipative Loaded Lines with Distributed Inductance

From the latter portion of Part III it will be recalled that the approximate effect of small distributed inductance is to alter slightly

the nominal impedance and the critical frequency of the loaded line without much affecting the relative impedance when expressed as a function of the relative frequency, over the first transmitting band and the lower part of the succeeding attenuating band. Thus an approximate way of taking account of the effects of small distributed inductance is to deal with the constants L_0' and C_0 of the corresponding "principal simulative loaded line"; since this line has no distributed inductance it is seen that the networks described in Part V for loaded lines without distributed inductance are adequate for loaded lines with small distributed inductance; the design-formulas remain unchanged beyond substituting L_0' for L' and C_0 for C; however, the simulative precision of the networks is altered slightly.

A slightly better approximation may be secured by working not only with L_0' and C_0 but also with fictitious values of σ and σ', say σ_0 and σ_0', slightly different from those which would be best if there were no distributed inductance.

Owing to the presence of a certain amount of distributed inductance in all transmission lines (even in cables), simulation of the σ'-load impedance ($\sigma' > \sigma_b'$) by means of a fractional-load (σ_b') type of basic network built out to σ'-load is slightly more precise than simulation of the σ-section impedance ($\sigma = \sigma'$) by means of a fractional-section (σ_b) type of basic network built out to σ-section. This is evident from the latter portion of Part III of this paper.

(Regarding the effects of small distributed inductance in loaded lines, Patent No. 1167693 may be of some interest.)

PART VII

Networks for Dissipative Loaded Lines

A natural first-approximation network for simulating the impedance of a dissipative loaded line is the network for the corresponding non-dissipative loaded line, the excess impedance thus being neglected; in the case of a high grade loaded line this is a good approximation except at very low frequencies. Various forms and types of networks for non-dissipative loaded lines having the basic relative terminations were described in Parts V and VI; those networks ("basic networks") can be built-out readily to any relative terminations by means of simple non-dissipative building-out structures.

When the excess impedance of the loaded line is not negligible an excess-simulator is required. A first-approximation excess-simulator for a loaded line is the excess-simulator for the corresponding

smooth line.) This is a good approximation over about the lower half or two-thirds of the transmitting band; but to be adequate in the upper part of the transmitting band it requires some modification in its proportioning or even in its form, according to several circumstances, such as the relative termination, the amount and distribution of the dissipation, and the ratio of the highest contemplated frequency to the critical frequency. The immediate neighborhood of the critical frequency is here disregarded, as having thus far been unimportant in practice; modification of the networks to extend their range of simulation right up to the critical frequency appears to present much greater difficulties.

PART VIII

Applications of the Simulating and the Compensating Networks

In this Part a considerable number of applications of the above-described networks will be outlined. (For some details and further applications, reference may be made to the patents cited in Part V—namely, Patent No. 1124901, No. 1167693, and No. 1437422, pertaining to the simulating networks; and No. 1243066 and No. 1475997 pertaining to the compensating networks.)

Applications of the Simulating Networks

Foremost of the uses of the simulating networks is their employment for balancing purposes in connection with 22-type repeaters, already spoken of in the Introduction.

Another application of a simulating network is for terminating an actual loaded line in the field or an artificial loaded line in the laboratory in such a way as to avoid reflection effects. For this purpose the proper terminating impedance is evidently one equal to the complementary characteristic impedance of the loaded line. Such a terminating impedance is often needed in the making of electrical tests or electrical measurements on a loaded line.

Furthermore, in making certain tests on apparatus normally associated with a loaded line, such line may be represented conveniently by the appropriate simulating network.

Applications of the Compensating Networks

The compensating networks have a wide variety of uses as neutralizing networks and also as simulating networks. These uses depend

mainly on the fact that a compensating network when used as a neutralizer enables the impedance of a loaded line to simulate approximately the impedance of a smooth line and hence to simulate at least roughly a constant resistance, and when used as a simulator enables the impedance of a smooth line to simulate approximately the impedance of a loaded line.

Foremost of the uses of the compensating networks is their employment for properly connecting together a loaded line and a smooth line, to reduce reflection effects at the junction. This may be accomplished either by means of the reactance compensator (Fig. 16) or by means of the susceptance compensator (Fig. 17) by adopting a suitable relative termination for the loaded line in each method. In describing these two methods, it will be assumed at first that the loaded line and the smooth line are non-dissipative and have equal nominal impedances. In the first method of compensation the loaded line is terminated at σ-section with σ in the neighborhood of 0.8, where its curve of characteristic resistance is nearly flat; and a reactance-compensator (Fig. 16) is inserted in series between the two lines. This compensator, by neutralizing the reactance of the given loaded line, makes that line appear like a smooth line; while, by simulating the complementary characteristic reactance of the loaded line, it makes the smooth line appear complementary to the given loaded line. In the second method of compensation the loaded line is terminated at σ'-load with σ' in the neighborhood of 0.8, where its curve of characteristic conductance is nearly flat; and a susceptance-compensator (Fig. 17) is inserted in shunt between the two lines at their junction. This compensator, by neutralizing the susceptance of the given loaded line, makes that like appear like a smooth line; while, by simulating the characteristic susceptance of the complementary loaded line, it makes the smooth line appear complementary to the given loaded line.

When, as actually, the lines are dissipative, the compensator continues to make the loaded line appear approximately like a smooth line, and to make the smooth line appear approximately like a loaded line; but now, unless the lines happen to be about equally dissipative, there will exist at their junction an irregularity arising chiefly from inequality in their "excess-impedances." This irregularity can be largely prevented from occurring when the gage of either or both of the lines is at the disposal of the designer; when this is not the case and the irregularity is seriously large, resort may be had to special equalizers termed "excess-impedance equalizers."

When the nominal impedances of the two lines are unequal, adjustment in that respect can be made by means of a transformer of suitable ratio.

Some other uses for the compensators are as follows: (a) to properly connect a loaded line to a repeater system whose impedance is nearly constant resistance; (b) to connect a loaded line type of filter (low-pass filter) to an amplifying element whose impedance is nearly constant resistance; (c) to connect a loaded line to terminal apparatus whose impedance is nearly constant resistance; (d) to convert the impedance of a loaded line to that of the corresponding smooth line and thereby enable it to be simulated (or to be balanced) by a smooth-line type of simulating network; (e) to convert the impedance of a smooth line to that of a loaded line and thereby enable it to be simulated (or to be balanced) by a loaded-line type of simulating network; (f) to neutralize the characteristic reactance of an approximately non-dissipative loaded line, thereby enabling the resulting nearly pure resistance impedance to be closely simulated (or to be closely balanced) by the network (Fig. 18) simulating the characteristic resistance of the loaded line; or—though somewhat less closely—by a mere resistance element; (g) to neutralize the characteristic susceptance of an approximately non-dissipative loaded line, thereby enabling the resulting nearly pure conductance admittance to be closely simulated (or to be closely balanced) by the network (Fig. 19) simulating the characteristic conductance of the loaded line; or—though somewhat less closely—by a mere conductance element.

In applications (a), (b), (c) the irregularity at the junction can be still further reduced by the addition of an excess simulator for simulating the excess impedance of the loaded line.

APPENDIX A

The Transmitting and the Attenuating Bands of a Non-Dissipative Loaded Line with Distributed Inductance

This Appendix contains the derivations of the formulas in Part III pertaining to the disposition of the transmitting and the attenuating bands; and also several alternative formulas; it outlines six graphical methods for studying the bands; and it discusses, more comprehensively than in the body of the paper, the salient properties of the bands and the effects produced by varying certain of the parameters.

Disposition of the Transmitting and the Attenuating Bands

The propagation constant $\Gamma = A + iB$ of a non-dissipative loaded line (per periodic interval) can be expressed in terms of $\lambda = L\ L'$ and the quantity D defined by equation (16). From Appendix D,

$$\cosh \Gamma = \cos 2D - \frac{D}{\lambda} \sin 2D, \tag{1-A}$$

$$\sinh^2 \Gamma = (\sin^2 2D)(D \tan D - \lambda)(D \cot D + \lambda)\ \lambda^2 \tag{2-A}$$

$$= (\sin^2 2D)(D^2 - \lambda^2 - 2\lambda D \cot 2D)\ \lambda^2 \tag{3-A}$$

$$= (-\sin^2 2D)(1 + 1\ \lambda) Z'_3. \tag{3.1-A}$$

Thus, for a non-dissipative loaded line, $\cosh \Gamma$ and $\sinh^2 \Gamma$ are both pure real.

When $\cosh \Gamma$ is known, A and B can be evaluated by means of the identity

$$\cosh \Gamma = \cosh (A + iB) = \cosh A \cos B + i \sinh A \sin B. \tag{4-A}$$

In particular, when $\cosh \Gamma$ is pure real—as for a non-dissipative loaded line—the values of A and B must evidently be such as to satisfy the pair of equations

$$\sinh A \sin B = 0, \qquad (5\text{-A}) \qquad \cosh A \cos B = \cosh \Gamma; \tag{6-A}$$

with, of course, the added restriction that A must be real and positive, and B real. Thence it is readily found that:

$$\begin{aligned} &\text{When } \cosh^2 \Gamma < 1, \text{ that is, } \sinh^2 \Gamma < 0, \\ &\text{then } A = 0 \text{ and } B = \cos^{-1} \cosh \Gamma; \end{aligned} \tag{7-A}$$

$$\begin{aligned} &\text{When } \cosh^2 \Gamma > 1, \text{ that is, } \sinh^2 \Gamma > 0, \\ &\text{then } A = \cosh^{-1} \cosh \Gamma \text{ and } B = q\pi; \end{aligned} \tag{8-A}$$

$\cosh \Gamma$ being real, and q being an even or an odd integer according as $\cosh \Gamma$ is positive or negative, respectively.

Before continuing with the general case ($\lambda \neq 0$) it seems worth while to digress long enough to apply the preceding general formulas to the limiting case where $\lambda = 0$. For it, formula (1-A) reduces to

$$\cosh \Gamma = 1 - 2r^2, \tag{9-A}$$

where $r = f\ f_c = D\ D_c$, and f_c is given by (3). Application of (7-A) and (8-A) to (9-A) shows that:

$$\text{When } 0 < r < 1, \text{ then } A = 0 \text{ and } B = 2 \sin^{-1} r; \tag{10-A}$$

$$\text{When } r > 1, \text{ than } A = 2 \cosh^{-1} r \text{ and } B = q\pi, \tag{11-A}$$

where q is an odd integer.

For illustrative purposes, Fig. 22 gives graphs of A and B throughout the first transmitting band ($0<r<1$) and part of the succeeding attenuating band, for a non-dissipative loaded line, with $\lambda=0$ and with $\lambda=0.12$. Of course, A is zero in the range $0<r<1$.

Returning now to the general case ($\lambda\neq 0$), we see that the transmitting bands ($A=0$) are characterized by the inequality $\sinh^2\Gamma<0$,

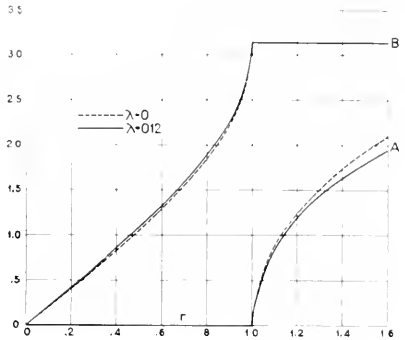


Fig. 22—Propagation Constant $\Gamma=A+iB$ in the First Transmitting Band ($0<r<1$) and in Part of the Succeeding Attenuating Band, of a Non-Dissipative Loaded Line with $\lambda=0$ and with $\lambda=0.12$

and the attenuating bands ($A\neq 0$) by the inequality $\sinh^2\Gamma>0$; and hence the transition points between the two kinds of bands are characterized by the equation $\sinh^2\Gamma=0$.

We seek the transition values of D, that is, the values of D where $\sinh^2\Gamma=0$; and we seek the transmitting and the attenuating ranges of D, that is, the ranges of D where $\sinh^2\Gamma<0$ and $\sinh^2\Gamma>0$, respectively.

The transition values of D are perhaps most readily found from the equation for $\sinh^2\Gamma$ when written in the form (2-A). They are the zeros of the first three factors in the right-hand member of that equation. The zeros of the factor $\sin^2 2D$ are at $D=m\pi/2$, with $m=0, 1, 2, 3, \ldots$; thus they subdivide the D-scale into segments of width $\pi/2$ each, as represented by Fig. 6; and they have the values represented by (18). The zeros of the factors $D\tan D-\lambda$ and $D\cot D+\lambda$ are situated in the odd and even numbered segments, respectively, because, λ is positive; there is one and only one zero in each

segment. Thus, if D_n denotes the zero of $\sinh^2\Gamma$ situated in the nth segment, then

$$(n-1)\frac{\pi}{2} < D_n < n\frac{\pi}{2}. \tag{12-A}$$

Either analytically or graphically it is readily seen that, when λ is small, D_n is only slightly greater than $(n-1)\pi/2$; it approaches that value as a limit when n approaches infinity, for all finite values of λ. The power series formula (21) for D_n is derived at a little later point in this Appendix.

Formulated analytically, with the arguments of the trigonometric functions reduced to the smallest positive values that preserve the values of the functions, the transition values of D are the values of $D_{n,n+1}$ and D_n satisfying the equations

$$\sin^2 2\left(D_{n,n+1} - n\frac{\pi}{2}\right) = 0, \tag{13-A}$$

$$D_n \tan\left(D_n - [n-1]\frac{\pi}{2}\right) = \lambda, \tag{14-A}$$

with $n = 0, 1, 2, 3, \ldots$ in (13-A) and $n = 1, 2, 3, \ldots$ in (14-A). Equation (13-A) is equivalent to $\sin^2 2D = 0$. With n odd and with n even, (14-A) is equivalent respectively to $D\tan D - \lambda = 0$ and to $D\cot D + \lambda = 0$. An equivalent of (14-A) is obtainable from the second factor of (3-A). By (3.1-A), still another equivalent is $Z'_{3} = 0$; that is, the values of D_n are the zeros of the mid-load relative impedance Z'_{3}, and hence of the mid-load impedance K'_{3}.

With $(n-1)\pi/2$ denoted by d_n, equation (14-A) shows that

$$D_n - d_n < \lambda \quad d_n, \quad (n = 2, 3, 4, \ldots) \qquad\qquad D_1 < \lambda.$$

By inspection of (2-A) it can be readily verified that $\sinh^2\Gamma$ is negative when $D_{n-1,n} < D < D_n$ and positive when $D_n < D < D_{n,n+1}$; and hence that these two ranges of D are a transmitting band and an attenuating band, respectively, the corresponding compound band thus being the range $D_{n-1,n} < D < D_{n,n+1}$. In this connection it may be of some academic interest to note that, strictly speaking, $D = 0$ is not a transition value of D between a transmitting and an attentuating band. For (2-A) shows that $\sinh^2\Gamma$ does not change sign when D passes through 0; on the contrary, $\sinh^2\Gamma$ is entirely unchanged when D is changed to $-D$. Thus, $D = 0$ is a point of symmetry, but not a transition point.

The values of D_n, namely, the roots of (14-A), cannot be written down directly or expressed exactly. But they can be found to any

desired degree of approximation by first developing the left side of (14-A) into a power series involving D_n; and then, by successive approximation or by undetermined coefficients, solving the resulting equation so as to express D_n as a power series in λ (that is, "reverting" the first series to obtain the second).

Digression on the Reversion of Power Series

Since there will be several occasions here for reverting a power series it seems worth while to digress sufficiently to furnish the requisite general formulas for the reversion of power series:[8]

Given $y = F(x)$ developed as a convergent power series in x,

$$y = x + a_2x^2 + a_3x^3 + a_4x^4 + \ldots . \qquad (15\text{-A})$$

The coefficient of x has been assumed to be unity because the formulation of the reversion is much simplified thereby without any real sacrifice of generality; for, if the coefficient of x were a_1, the equation could be reduced immediately to the form (15-A), either by treating a_1x as the independent variable, or by dividing through by a_1 and then treating y a_1 as the dependent variable.

The given equation (15-A) expresses y as a power series in x. It is required to revert this relation, that is, to express x as a power series in y. In the present work this was done originally by successive approximation, and was verified later by the method of undetermined coefficients. Evidently the first approximation to the solution of (15-A) is merely $x_1 = y$, and thence the second approximation is $x_2 = y - a_2x_1^2 = y - a_2y^2$. But the higher approximations cannot be written down thus directly; indeed the labor of obtaining them increases rapidly. The work was carried through the sixth approximation, with the result:

$$x = y + (-a_2)y^2 + (2a_2^2 - a_3)y^3 + (-5a_2^3 + 5a_2a_3 - a_4)y^4$$
$$+ (14a_2^4 - 21a_2^2a_3 + 6a_2a_4 + 3a_3^2 - a_5)y^5$$
$$+ (-42a_2^5 + 84a_2^3a_3 - 28a_2^2a_4 - 28a_2a_3^2 + 7a_2a_5 + 7a_3a_4 - a_6)y^6 + \ldots . \qquad (16\text{-A})$$

[8] Cf., for instance, Bromwich, "Theory of Infinite Series"; Goursat-Hedrick, "Mathematical Analysis"; Wilson, "Advanced Calculus"; Chrystal, "Text Book of Algebra." But in none of these references is the reversion carried far enough; moreover, the formulas there obtained do not apply directly to a series containing only even powers—one of the cases in the present application. At considerable labor, by two independent methods, I remedied both of these lacks. Somewhat later I came upon a valuable article by C. E. Van Orstrand, "The Reversion of Power Series" (*Phil. Mag.*, March, 1910), where the reversion is carried to no less than thirteen terms, but is not directly applicable to series containing only even powers.

This was verified by the method of undetermined coefficients, consisting in assuming

$$x = y + b_2y^2 + b_3y^3 + b_4y^4 + \ldots$$

and then substituting this expression for x into (15-A) to evaluate the b's by treating the resulting equation as an identity.

In the degenerate case where only even powers of x are present in (15-A) the formula (16-A) when applied directly does not correctly express the solution (for reasons appearing below). However, the given equation, containing only even powers of x, say

$$y = x^2 + c_2x^4 + c_3x^6 + c_4x^8 + \ldots, \qquad (17\text{-A})$$

can be correctly solved for (x^2) by direct application of (16-A), with $a_s = c_s$; and then the value of x can be expressed as a power series in y by extracting the square root of the power series representing (x^2). In that way the solution of (17-A) was found to be

$$\frac{x}{\sqrt{y}} = 1 + \left(-\frac{1}{2}c_2\right)y + \left(\frac{7}{8}c_2^2 - \frac{1}{2}c_3\right)y^2 + \left(-\frac{33}{16}c_2^3 + \frac{9}{4}c_2c_3 - \frac{1}{2}c_4\right)y^3$$
$$+ \left(\frac{715}{128}c_2^4 - \frac{143}{16}c_2^2c_3 + \frac{11}{4}c_2c_4 + \frac{11}{8}c_3^2 - \frac{1}{2}c_5\right)y^4 + \left(-\frac{4199}{256}c_2^5\right.$$
$$\left. + \frac{1105}{32}c_2^3c_3 - \frac{195}{16}c_2^2c_4 - \frac{195}{16}c_2c_3^2 + \frac{13}{4}c_2c_5 + \frac{13}{4}c_3c_4 - \frac{1}{2}c_6\right)y^5 + \ldots. \qquad (18\text{-A})$$

This result was verified by the method of undetermined coefficients, by writing x in the form

$$x = \sqrt{y}\,(1 + e_1y + e_2y^2 + e_3y^3 + \ldots) \qquad (18.1\text{-A})$$

and then evaluating the e's by substituting (18.1-A) into (17-A). Still another method would be to extract the square root of (17-A) as the first step, thereby expressing $\sqrt{y}$ as a power series in x of the form (15-A); and then reverting by application of (16-A), thereby expressing x as a power series in $\sqrt{y}$ and thence of the form (18.1-A).

For use in this connection it may be noted that the square root of a power series having the form

$$y^2 = 1 + h_1x + h_2x^2 + h_3x^3 + \ldots$$

will be of the form

$$y = 1 + k_1x + k_2x^2 + k_3x^3 + \ldots.$$

The k's can be evaluated by identifying the first equation with the square of the second; their values are found to be

$$k_1 = \tfrac{1}{2}h_1, \qquad k_2 = \tfrac{1}{2}h_2 - \tfrac{1}{2}k_1^2, \qquad k_3 = \tfrac{1}{2}h_3 - k_1k_2,$$
$$k_4 = \tfrac{1}{2}h_4 - \tfrac{1}{2}k_2^2 - k_1k_3, \qquad k_5 = \tfrac{1}{2}h_5 - k_1k_4 - k_2k_3,$$
$$k_6 = \tfrac{1}{2}h_6 - \tfrac{1}{2}k_3^2 - k_1k_5 - k_2k_4.$$

Derivations of Formulas for the Transition Points

The above general formulas for the reversion of power series will now be applied in the derivation of the formulas (21) and (22) for D_n and D_1, in the body of the paper; and also in the derivation of certain other formulas, not included there.

To outline the derivation of the formula (21) for D_n, denote $(n-1)\pi$ 2 by d_n and $D_n - d_n$ by τ_n, so that (14-A) becomes

$$(d_n + \tau_n) \tan \tau_n = \lambda. \tag{19-A}$$

Now replace $\tan \tau_n$ by its known power series expression, and divide both sides of the resulting equation by d_n; thus (19-A) becomes

$$\frac{\lambda}{d_n} = \tau_n + \frac{1}{d_n}\tau_n^2 + \frac{1}{3}\tau_n^3 + \frac{1}{3d_n}\tau_n^4 + \frac{2}{15}\tau_n^5 + \frac{2}{15d_n}\tau_n^6 + \ldots . \tag{20-A}$$

This is of the form (15-A), and hence can be reverted by direct application of (16-A); the result is (21).

An alternative formula for D_n can be obtained by starting from Gregory's series,

$$v = \tan v - \frac{\tan^3 v}{3} + \frac{\tan^5 v}{5} - \frac{\tan^7 v}{7} + \ldots . \tag{20.1-A}$$

Application of this to (19-A) enables the left side of that equation to be expressed as a power series in $\tan \tau_n$; and when the resulting equation is reverted by means of (16-A) and then τ_n replaced by $D_n - d_n$ the result is

$$\tan (D_n - d_n) = \frac{\lambda}{d_n} - \frac{1}{d_n}\left(\frac{\lambda}{d_n}\right)^2 + \frac{2}{d_n^2}\left(\frac{\lambda}{d_n}\right)^3 - \left(\frac{5}{d_n^3} - \frac{1}{3d_n}\right)\left(\frac{\lambda}{d_n}\right)^4$$
$$+ \left(\frac{14}{d_n^4} - \frac{2}{d_n^2}\right)\left(\frac{\lambda}{d_n}\right)^5 - \left(\frac{42}{d_n^5} - \frac{28}{3d_n^3} + \frac{1}{5d_n}\right)\left(\frac{\lambda}{d_n}\right)^6 + \ldots . \tag{20.2-A}$$

It has already been noted that (21) is not valid for $n = 1$ and hence does not include the formula (22) for D_1. To obtain this formula for D_1, start with the equation

$$D_1 \tan D_1 = \lambda, \tag{21-A}$$

obtained by setting $n = 1$ in (14-A). Then replace $\tan D_1$ by its known power series expansion, thus obtaining the equation

$$\lambda = D_1^2 + \frac{1}{3}D_1^4 + \frac{2}{15}D_1^6 + \frac{17}{315}D_1^8 + \frac{62}{2835}D_1^{10} + \frac{1382}{155925}D_1^{12} + \ldots . \tag{22-A}$$

This is of the form (17-A), and hence can be reverted by direct application of (18-A); the result is (22).

It may be noted that (22-A), when regarded as a power series in (D_1^2), is of the form (15-A) and hence that (D_1^2) can be expressed as a power series in λ by direct application of (16-A); the result is[7]

$$D_1^2 = \lambda - \frac{\lambda^2}{3} + \frac{13\lambda^3}{45} - \frac{16\lambda^4}{945} + \frac{16\lambda^5}{14175} + \frac{64\lambda^6}{93555} \cdots \quad (23\text{-A})$$

In certain applications this formula for D_1^2 is more useful than formula (22) for D_1; though the two are ultimately equivalent. A formula for p^2 is obtainable by dividing both sides of (23-A) by λ; for $p^2 = D_1^2\ \lambda$, by (16).

An alternative formula for D_1 can be obtained by starting from Gregory's series (20.1-A). Application of this to (21-A) enables the left side of that equation to be expressed as a power series in $\tan D_1$; and when the resulting equation is reverted by means of (18-A) the result is[7]

$$\tan D_1 = \lambda\ \lambda\left(1 + \frac{\lambda}{6} - \frac{\lambda^2}{360} - \frac{11\lambda^3}{5040} + \frac{1357\lambda^4}{1814400} \cdots\right). \quad (23.1\text{-A})$$

Series that are even more convergent than (21) and (22), though much less simple, can be obtained by expanding the original function in the neighborhood of a value of the variable known to be an approximate solution of the equation to be solved, and then reverting the resulting series. To formulate the procedure analytically and generally, let u denote the variable, and $\psi(u)$ the function; and let the equation to be solved for u be

$$\psi(u) = q. \quad (24\text{-A})$$

Then, if U is an approximate solution of this equation, application of Taylor's theorem leads to the following implicit equation for $u - U$:

$$\frac{q - \psi(U)}{\psi'(U)} = (u - U) + \frac{(u-U)^2}{2!}\frac{\psi''(U)}{\psi'(U)} + \frac{(u-U)^3}{3!}\frac{\psi'''(U)}{\psi'(U)} + \cdots \quad (25\text{-A})$$

The left side of this is known. The right side is a power series in $u - U$, with U known; the better the approximation represented by U, the more rapidly convergent is the series. This equation (25-A) in $u - U$ is of the form (15-A), with

$$y = \frac{q - \psi(U)}{\psi'(U)}, \qquad x = u - U, \qquad a_s = \frac{\psi^{(s)}(U)}{s!\,\psi'(U)}; \quad (26\text{-A})$$

and thence (25-A) can be reverted by application of (16-A), so that $u - U$ will be expressed as a power series in $[q - \psi(U)]\ \psi'(U)$.

To apply the above general method in order to obtain for D_n a series more convergent than (21), return to (19-A) and note that when λ is small a first approximation for τ_n is $\tau_n = \lambda\ d_n$. Then apply (16-A), with y, x, and a_s having the values expressed by (26-A); and $q = \lambda$, $u = \tau_n$, $U = \lambda\ d_n$, and $\psi(u) = (u + d_n) \tan u$. The formulas for the first few successive derivatives of $\psi(u)$ will be needed, of course.

Similarly, to obtain for D_1 a series more convergent than (22), return to (21-A) and note that when λ is small a first approximation for D_1 is $D_1 = \sqrt{\lambda}$. Then apply (16-A), with y, x, and a_s having the values expressed by (26-A); and $q = \lambda$, $u = D_1$, $U = \sqrt{\lambda}$, and $\psi(u) = u \tan u$.

Graphical Methods for Locating the Tansition Points

The positions of the transition points D_n ($n = 1, 2, 3, \ldots$) on the D-scale can be determined also graphically, in several different ways corresponding to several different ways of writing the function $(D \tan D - \lambda)\ (D \cot D + \lambda)$ whose zeros are the values of D_n. To formulate such graphical methods concisely, let E denote any function of the variable D, so that, geometrically, E is the ordinate corresponding to the abscissa D. Six of the various possible graphical methods are then briefly but completely indicated by the following respective statements that the points D_n are the abscissas of the points of intersection of:

1. The horizontal straight line $E = \lambda$ with the curves $E = D \tan D$; the horizontal straight line $E = -\lambda$ with the curves $E = D \cot D$.
2. The straight line $E = D$ with the curves $E = \lambda \cot D$; the straight line $E = -D$ with the curves $E = \lambda \tan D$.
3. The straight line $E = D\ \lambda$ with the cotangent curves $E = \cot D$; the straight line $E = -D\ \lambda$ with the tangent curves $E = \tan D$.
4. The hyperbola $E = \lambda\ D$ with the tangent curves $E = \tan D$; the hyperbola $E = -\lambda\ D$ with the cotangent curves $E = \cot D$.
5. The parabola $E = D^2\ \lambda - \lambda$ with the curves $E = 2D \cot 2D$.
6. The curve $E = D\ 2\lambda - \lambda\ 2D$, compounded of the straight line $E = D\ 2\lambda$ and the hyperbola $E = -\lambda\ 2D$, with the cotangent curves $E = \cot 2D$.

In methods 1, 2, 3, 4, the first set of intersections is situated in the odd-numbered segments, the second set in the even numbered segments; each segment of width $\pi/2$.

Besides being susceptible of quantitative service, these graphical methods are useful for qualitative purposes. For instance, they show

clearly that: one and only one transition value of D lies within each segment of width $\pi/2$; $\sinh^2\Gamma<0$ when $D_{n-1,n}<D<D_{n,}$ and $\sinh^2\Gamma>0$ when $D_n<D<D_{n,n+1}$; the zeros of $\lambda-D\tan D$ and of $\lambda+D\cot D$ are situated in the odd and even numbered segments, respectively; with increasing D, the transmitting bands continually decrease in width and the attenuating bands continually increase in width, the change taking place rapidly at first and then more and more slowly; the mid-point relative impedances are pure imaginary throughout every attenuating band and pure real throughout every transmitting band, and, they have the ranges stated in the third and fourth paragraphs following equation (26.1). The graphical methods are useful also for showing the nature of the effects produced by varying the parameter λ.

Discussion of the Disposition of the Bands

The rest of this Appendix will be devoted to a discussion of the most salient properties of the compound bands and their constituent transmitting and attenuating bands.

The ratio of transmitting band width to compound band width continually decreases with increasing D and becomes zero when D becomes infinite; that is, the transmitting bands vanish and the compound bands become pure attenuating bands. These facts can be seen graphically, or analytically from equation (14-A).

The ratio of transmitting band width to compound band width continually increases with increasing λ; this ratio ranging from zero when λ is zero to unity when λ is infinite. These facts can be seen graphically, or from equation (14-A). When λ approaches zero the f-width of each compound band approaches infinity; the f-width of each transmitting band approaches zero, except for the first transmitting band, whose width approaches a value equal to $f'_1=f'_c$—for equation (14-A) shows that $D_n(D_n-D_{n-1,n})\lambda$ approaches unity, and hence that $f_n(f_n-f_{n-1,n})$ approaches $\frac{1}{4}\pi^2L'C=f'^2_1$, whence $f_n-f_{n-1,n}$ approaches zero for $n\neq 1$ and approaches f'_1 for $n=1$.

The effects of varying the parameter λ will now be outlined briefly, in the next two paragraphs, for the cases respectively of $L'C$ fixed and LC fixed. The conclusions reached depend partly on the equation $D=\frac{1}{2}\omega\sqrt{LC}=\frac{1}{2}\omega\sqrt{\lambda L'C}$ defining D; partly on the fact already deduced that the D-width of each compound band is an absolute constant ($\pi/2$); and partly on equation (14-A).

When $L'C$ is fixed, increasing λ reduces all of the transition frequencies. The transition frequencies bounding the compound bands,

and hence the widths of the compound bands, decrease in direct proportion to increase of $\sqrt{\lambda}$. The internal transition frequencies, however, do not decrease so rapidly; for the ratio of transmitting band width to attenuating band width increases with increasing λ. When λ approaches infinity each compound band approaches a width of zero, but the ratio of transmitting band width to compound band width approaches unity; so that when λ becomes infinite there are within any finite frequency range an infinite number of compound bands which are pure transmitting bands. On the other hand, when λ approaches zero the compound bands approach infinite width and hence move out toward infinity, except that the left end-point of the first band is fixed at $f=0$. When λ has become zero the first compound band has expanded to an infinite width; and its critical value f_1 of f has become equal to the limiting value $f'_1=1/\pi\sqrt{L'C}$ as can be seen from (14-A) by putting $n=1$ and then applying the relation $D\sqrt{\lambda}=\frac{1}{2}\omega\sqrt{L'C}$.

When LC is fixed the f-widths and locations of the compound bands are independent of λ, but the widths of the constituent attenuating and transmitting bands depend on λ; that is, the boundary points $f_{n-1\,n}$ and $f_{n,n+1}$ of the nth compound band are independent of λ, but the internal transition point f_n depends on λ. With increasing λ the attenuating bands become continually narrower, and vanish when λ becomes infinite, the transmitting bands thereby coalescing to form a pure transmitting band extending from zero to infinity. With decreasing λ the transmitting bands become continually narrower, and vanish when λ becomes zero, the attenuating bands thereby coalescing to form a pure attenuating band extending from zero to infinity.

APPENDIX B

Theoretical Bases of the Simulating Networks in Figs. 12 and 13

The Impedance-Simulator in Fig. 12

This network takes advantage of the fact, depicted in Fig. 5, that the graph of the σ-section characteristic resistance of a loaded line, for values of σ in the neighborhood of 0.2, is nearly flat over most of the transmitting band and hence can be approximately simulated by a mere constant resistance chosen approximately equal to the nominal impedance $\sqrt{L'/C}$. This is the basis for the R_1-portion of the network in Fig. 12. The basis for the L_1C_1-portion is the fact (proved in Appendix C) that, in the transmitting band, the σ-section

characteristic reactance can be exactly simulated (for any fixed value of σ between 0 and 1.2) by the network in Fig. 16.

The Admittance-Simulator in Fig. 13

This network takes advantage of the fact, depicted in Fig. 5, that the graph of the σ'-load characteristic conductance of a loaded line, for values of σ' in the neighborhood of 0.2, is nearly flat over most of the transmitting band and hence can be approximately simulated by a mere constant conductance chosen approximately equal to the nominal admittance $\sqrt{C/L'}$. This is the basis for the G_1'-portion of the network in Fig. 13. The basis for the $L'_1C''_1$-portion is the fact (proved in Appendix C) that, in the transmitting band, the σ'-load characteristic susceptance can be exactly simulated (for any fixed value of σ' between 0 and 1.2) by the network in Fig. 17.

APPENDIX C

Derivations of the Design-Formulas for the Compensating Networks in Figs. 16 and 17

The Reactance-Compensator in Fig. 16

For any values of C_5 and L_5 the reactance T of this network is

$$T = \frac{\omega L_5}{1 - \omega^2 L_5 C_5}.$$

By equation (4) the characteristic reactance N of the loaded line within its transmitting band is

$$N = \frac{k(1-2\sigma)\omega\ \omega_c}{1 - 4\sigma(1-\sigma)\omega^2\ \omega_c^2}.$$

Comparison of these two equations shows that T and N are of the same functional form in ω; and that the conditions for T to be identically equal to $\pm N$ are

$$L_5 = \pm k(1-2\sigma)\ \omega_c, \qquad L_5 C_5 = 4\sigma(1-\sigma)\ \omega_c^2,$$

whence

$$C_5 = \pm 4\sigma(1-\sigma)\ (1-2\sigma)k\omega_c,$$

the upper and the lower sign of $\pm$ corresponding to the use of the compensator as a reactance-simulator and a reactance-neutralizer, respectively. These values of L_5 and C_5 are equivalent to those appearing in Fig. 16, because $k = \sqrt{L'/C}$ and $\omega_c = 2\pi f_c = 2\sqrt{L'C}$.

For positive values of L_5 the equation for L_5 shows that $\sigma \gtrless 1/2$, corresponding to $\pm$; and then the equation for C_5 shows that $\sigma \gtrless 0$, corresponding to $\pm$. Hence $0<\sigma<1/2$ for $T=+N$, and $1/2<\sigma<1$ for $T=-N$.

The Susceptance-Compensator in Fig. 17

For any values of C_5' and L_5' the susceptance S' of this network is

$$S'=\frac{\omega C_5'}{1-\omega^2 L_5' C_5'}.$$

By equation (5) the characteristic susceptance Q' of the loaded line within its transmitting band is

$$Q'=\frac{h(1-2\sigma')\omega\ \omega_c}{1-4\sigma'(1-\sigma')\omega^2\ \omega_c^2}.$$

Thus S' and Q' are of the same functional form in ω; and the conditions for S' to be identically equal to $\pm Q'$ are that

$$C_5'=\pm h(1-2\sigma')\ \omega_c,$$

$$L_5'=\pm 4\sigma'(1-\sigma')\ (1-2\sigma')h\omega_c,$$

the upper and the lower sign of $\pm$ corresponding to the use of the compensator as a susceptance-simulator and a susceptance-neutralizer respectively. These values of C_5' and L_5' are equivalent to those appearing in Fig. 17, because $h=\sqrt{C\ L'}$ and $\omega_c=2\ \sqrt{L'C}$.

The equations for C_5' and L_5' show that $0<\sigma'<1/2$ for $S'=+Q'$, and that $1/2<\sigma'<1$ for $S'=-Q'$.

APPENDIX D

General Formulas for the Characteristic Impedances and the Propagation Constant of Loaded Lines

For reference purposes this Appendix gives the general formulas for the mid-section ($\sigma=0.5$) and mid-load ($\sigma'=0.5$) characteristic impedances K_5 and K'_5 and the propagation constant Γ of a periodically loaded line (of the series type).

The symbols have the following meanings: d denotes the impedance of each load. g and γ pertain to the line before loading; g denotes the characteristic impedance, and γ denotes the propagation constant of a segment whose length is equal to the distance between adjacent loads after the line is loaded.

The formulas for the mid-section and mid-load characteristic impedances K_5 and K'_5 are[2]

$$K_5 = g\sqrt{\frac{1+\frac{d}{2g}\coth\frac{\gamma}{2}}{1+\frac{d}{2g}\tanh\frac{\gamma}{2}}}, \tag{1-D}$$

$$K'_5 = g\sqrt{\left(1+\frac{d}{2g}\coth\frac{\gamma}{2}\right)\left(1+\frac{d}{2g}\tanh\frac{\gamma}{2}\right)} \tag{2-D}$$

$$= g\sqrt{1+\frac{1}{4}\left(\frac{d}{g}\right)^2+\frac{d}{g}\coth\gamma}. \tag{3-D}$$

Several mutually equivalent formulas for the propagation constant Γ (per periodic interval) are:

$$\cosh\Gamma = \cosh\gamma + \frac{d}{2g}\sinh\gamma. \tag{4-D}$$

$$\sinh\Gamma = \frac{K'_5}{g}\sinh\gamma. \tag{5-D}$$

$$\tanh\tfrac{1}{2}\Gamma = \frac{K_5}{g}\tanh\tfrac{1}{2}\gamma. \tag{6-D}$$

The sending-end impedance J of any smooth line, of characteristic impedance g_1 and total propagation constant γ_1, whose distant end is closed through any impedance J_1, has the formula

$$J = g_1\frac{J_1\ g_1+\tanh\gamma_1}{1+(J_1\ g_1)\tanh\gamma_1}. \tag{7-D}$$

This enables the formula for the σ-section characteristic impedance K_σ of a loaded line to be established by starting with the formula (4-D) for the mid-section characteristic impedance K_5.

[2] Formulas (2-D) and (3-D) for K'_5 and formula (4-D) for $\cosh\Gamma$ are given by G. A. Campbell in his paper on loaded lines (*Phil. Mag.*, March, 1903 cited in footnote 2.

Some Contemporary Advances in Physics—IV

By KARL K. DARROW

Closing the Spectrum Gap Between the Infra-red and the Hertzian Regions

AN electrical circuit having a natural oscillation-frequency anywhere below 10^8 can be constructed by anyone with suitable condensers, inductance-coils, and a few feet of wire at his disposal. It can be set into oscillation by abruptly closing it when the condenser is charged, by coupling it to an audion, or otherwise; and the waves which it radiates while oscillating can be detected and measured, at least when the frequency exceeds 10^4. Thus it is possible to generate perceptible electromagnetic waves of frequencies up to 10^8, and hence of wavelengths down to 3 metres, by methods that may be called *electrotechnical*. Waves shorter than 300 cm., frequencies higher than 10^8 cycles, are not easily produced by any such method; for if one uses excessively small condensers and inductance-coils in the hope of forcing the circuit-frequency much past 10^8, or even omits coils and condensers altogether, it is found that the auxiliary apparatus, the audion, even the wires of the circuit themselves, possess capacities and inductances which can not be annulled and which hold the oscillation-frequency down. By devising oscillating systems which have scarcely any outward resemblance to the circuits of familiar experience (although a formal analogy can be established) Hertz and his successors generated electromagnetic waves of frequencies up to 10^{11} and wavelengths down to 3 mm. Beyond a certain gap there commences, near frequency 10^{12} and wavelength 0.3 mm., the far-flung spectrum of rays emitted by molecules and atoms. This interval is one of the two lacunae in the complete electromagnetic spectrum extending from 10^4 past 10^{20} cycles, which were mentioned in the preceding article of this series. Unlike the gap between the ultra-violet and the X-rays, it is not believed to be populated by rays resulting from important processes occurring within the atoms, nor do we know of any other peculiar type of radiation which should be sought within it; and perhaps the bridging of it, when finally and unquestionably achieved, will be held notable chiefly as a feat of experimental technique or a *tour de force*. On the other hand, so long as the gap remains unspanned, we can hardly dismiss the possibility that something in the order of nature may reserve one range of wavelengths for the "natural" rays resulting from atomic processes, and limit the "artificial" waves generable by electrotechnical

means to a distant range which never can be extended to overlap the other.

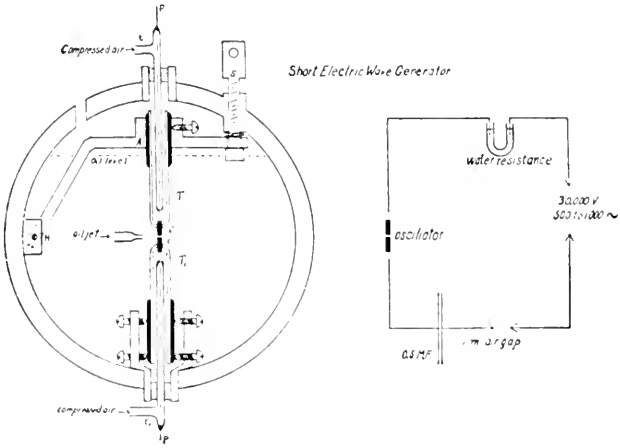
The advance into the lacuna from the direction of shorter waves, that is, from the spectrum of natural rays, came almost to a stop in 1911, at a wavelength between 0.3 and 0.4 mm. Rubens and von Baeyer examined the rays emitted by a mercury vapor arc in a quartz tube, operated with a comparatively high expenditure of power; they filtered the radiation through a succession of diaphragms and lenses which cut out a large fraction of the short-wave radiation, but not by any means all of it. At first they analyzed the radiation which came through with an interferometer, like the one which I shall describe in speaking of short artificial waves; the curves indicated that it consisted largely of two waves, one at 0.218 mm. and the other at 0.343 mm. Rubens in 1921 returned to the experiments, and diffracted the transmitted rays with a large-scale wire grating (the wires were a millimetre thick and a millimetre apart). This method of analyzing the radiation, in which it is spread out in a spectrum, is preferable to the other. The results were quite concordant with the earlier ones; the curves of intensity versus wavelength show maxima at 0.210 mm. and 0.325 mm., and extend out as far as 0.4 mm.[1] There is no sign that this is a definite physical limit; it is merely the point at which the rays become too feeble to produce an unmistakable deflection of the micro-radiometer. Nichols and Tear also have observed these long natural waves.

To advance into the lacuna from the region of artificial waves, it was found necessary first of all to remodel the oscillator or "doublet" by means of which Hertz had generated the first waves of this kind. The original oscillators of Hertz were rather large; some for example, consisted of pairs of metal plates 40 cm. square or pairs of spheres 30 cm. in diameter with arms projecting from each toward the other and carrying knobs several mm. or cm. in diameter; their natural frequencies were of the order $10^8 - 10^9$. Their successors were made progressively smaller, and the latest oscillators are comparatively minute—in dealing with a less exact science, one would describe them as microscopic; for Möbius before the war used a doublet of

[1] It is not necessarily to be assumed that the mercury arc is unique in sending out rays of so great a wavelength with so great an intensity; these rays may not be more intense than a black body of the same temperature as the arc would emit in this portion of its spectrum (although this interpretation would involve a rather high estimate of the arc temperature, many thousands of degrees). But if we had a black body of this temperature available, we might not be able to detect these rays because of the flood of light of higher frequencies which could not be completely deflected from the path of the long waves. Thus we are led to the paradoxical conclusion that the mercury arc may be unique not in furnishing these rays, but in not emitting so much radiation of lesser wavelengths that the rays desired cannot be isolated.

platinum cylinders each 1 mm. long [2] and 0.5 mm. thick, while Nichols and Tear in 1922 succeeded in making and using tungsten cylinders 0.2 mm. long and 0.2 mm. thick. To appreciate this feat it is necessary to realize that the cylinders must be sealed into a sheet of glass *with both ends projecting*; as they are shown in Fig. 1, which like the remaining figures (unless otherwise mentioned) comes from the work of Nichols and Tear.

In Fig. 1, the oscillator-cylinders are shown at c and c_1; they are sealed into the tips of hard-glass tubes T and T_1, and project outwards



Sketch of oscillator.

Fig. 1—Diagrams of the Oscillator and the Circuit Used by Nichols and Tear. (*Physical Review*)

into kerosene oil which fills the entire cylindrical container up to the level indicated by the dashed line.[3] The oscillator is excited by the voltage-impulses in the secondary of an induction-coil, resulting from

[2] The figure given by Möbius is 1.98 mm. (last column on p. 317, *l.c.infra*) which he says (on p. 302) applies to the *Gesamtlange* of the doublet. Theory indicates that the wavelength of the fundamental oscillation is about twice the length of the cylinders, but the exact value of the factor is in doubt.

[3] The kerosene, the "oil-jet" for keeping it circulating rapidly through the region between the cylinders and the blasts of compressed air into the tubes T and T_1 (note the spark-gaps in the leading in wires in these tubes) are all empirical devices for improving the efficiency of the apparatus.

abrupt breaks of the primary circuit produced by a mechanical interrupter at the rate of a thousand per second. Each of these voltage-impulses excites a spark between the doublet-cylinders, accompanied by a highly-damped oscillation which radiates what the authors describe as "a very short wave-train with from 60 to 80% of the energy concentrated in the first half-wavelength." This high damping is deplorable, as the waves are inconvenient to measure and must be regarded as mixtures of sine-waves of different frequencies. The gap between the cylinders is of the order 0.01–0.02 mm.; it changes rapidly and irregularly as the opposing surfaces are eaten away by

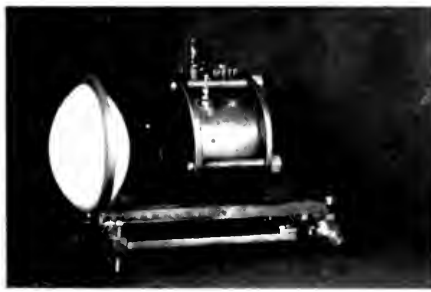


Fig. 2—Photograph of the Oscillator Used by Nichols and Tear. (*Physical Review*)

the sparks (tungsten was chosen by Nichols and Tear instead of platinum in the hope, justified by the event, of diminishing this trouble).

The rays issue through a mica window in the front of the containing-cylinder and are formed into a plane-parallel beam by an enormous double-convex paraffin lens (these objects are shown in the photograph, Fig. 2). Paraboloidal mirrors can be and have been used instead of the lens. In the sketch of Fig. 3, L_1 represents the lens; the plane-parallel beam proceeds to the mirror A and thence to the mirror B, which is really the pair of mirrors on the left-hand end of the apparatus of which Fig. 4 is a photograph. In this apparatus, the "Boltzmann interferometer," the upper mirror slides backward and forward (left to right and right to left, in the picture) along the guides, controlled by the screw; it remains always parallel to the lower and stationary mirror. Half of the plane-parallel beam falls upon each mirror, and the two reflected halves travel side by side

to the lens L_2 which merges them in a common focus at M, where the receiver stands.[4]

The intensity at M depends, by virtue of the principle of interference of periodic waves in its simplest conceivable application, on the ratio of the distance between the planes of the two mirrors to the

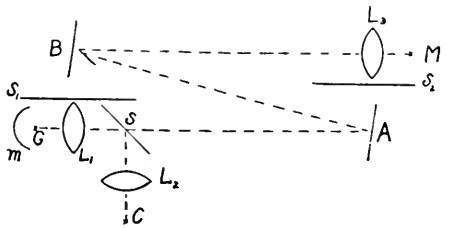


Diagram for wave-length measurements.

Fig. 3—Path of the Radiation from Oscillator to Receiver. (*Physical Review*)

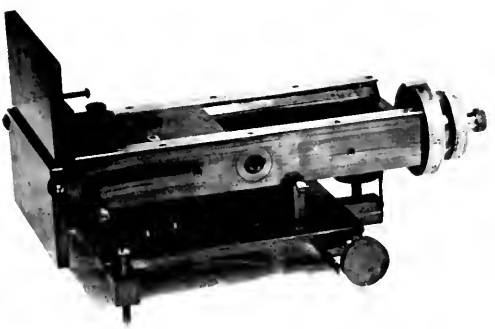


Fig. 4—Photograph of the Boltzmann Interferometer. (*Physical Review*)

wavelength of the rays. *If* at M there were a receiver of which the reading was perfectly proportional to the amplitude of the vibration at M and *if* the original wave-train were perfectly sinusoidal and

[4] In the sketch S is a semi-transparent mirror (glass ebonite, or cardboard) which reflects a part of the beam to a lens L_2 and focus C where its intensity can be measured at the same moment as the intensity at M. The variability of the output of the source makes this control indispensable

very long, then the curve obtained by displacing the movable mirror step by step and plotting the receiver-reading against the mirror-displacement would be a perfect sine-curve; the distance between the positions of the mirror corresponding to two consecutive maxima of the curve would be half the wavelength of the wave-train. Unfortunately neither the receiver nor the wave-train is ever perfect. The wave-train is a heavily-damped sinusoid, and consequently the curve of receiver-reading versus mirror-displacement flattens out before the mirror has been moved very far. Even so, the interpretation might not be uncertain if the receiver gave a reading proportional to the time integral of the intensity of the wave-motion at M. This it rarely does.

The receiver, in this region of the spectrum, must be a thermal receiver—a short thin wire or a narrow band of sputtered metal upon a strip of insulating substance, or sometimes a wire loop. In this the

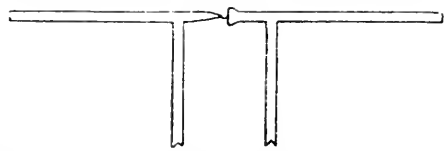


Fig. 5.—Thermoelectric Receiver Used by Mobius. (*Annalen der Physik*)

incident waves induce a resonance-current, of which the Joule heat produces the directly-measured effect. A thermojunction may be intercalated in the resonant wire, as in Möbius' apparatus (Fig. 5; in the middle of the transverse piece, 14 mm. long and 0.3 mm. thick, a platinum tip is welded into a tellurium socket). Nichols and Tear, developing a method introduced by G. F. Hull, mounted the thin wire or the sputtered ribbon in front of a radiometer-vane; the Joule heat warmed the front face of the vane, and the rather mysterious agencies sometimes called "radiometer forces" came into play. Four of their receivers are shown in Fig. 6 at *b*, *c*, *d*, and *e*. In each of these sketches V_1 represents the edge of the radiometer vane; *c* in sketch *b* is a wire running from end to end of it, while c_1, c_2, etc., in sketches *d* and *e* are short wires mounted vertically or horizontally behind it. The mounting is shown in Fig. 6*a*; the vanes are seen front-face, one having its wire or wires in front and the other behind, so that the radiometer forces on both will produce torques acting in the same sense. The vanes with the cross-pieces c_1 and c_2 are mounted upon the rod *q*, which is suspended from a torsion-fibre;

and from the rod is suspended a mirror m to indicate the amount of twist. The air-pressure is adjusted to produce the maximum torque.

The outstanding defect of a receiver of this type is, that it imprints its own characteristics upon the data. It will not respond effectively to a wave-train not possessing a frequency agreeing closely with its

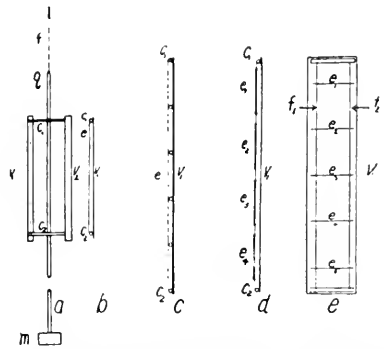


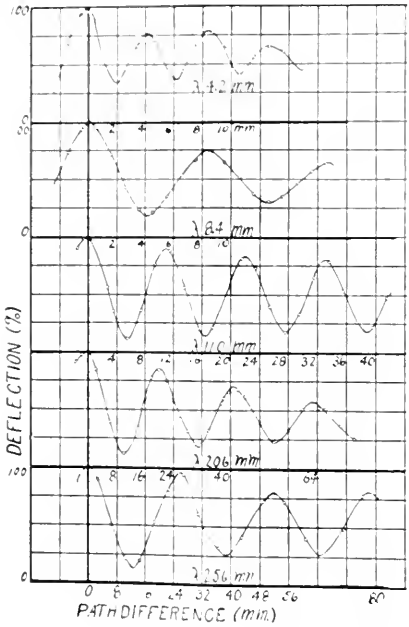
Fig. 6—Radiometer and Radiometer Vanes Used by Nichols and Tear. (*Physical Review*)

own, or with some harmonic of its own; and has a tendency to exaggerate the apparent proportion of such frequencies in a beam which is a mixture of frequencies, as a damped wave-train is. In general, the curve of receiver-reading versus mirror-displacement is an unevenly wavy one which, when analyzed into components in Fourier's manner, is found to contain at least two frequencies, one of which is attributed to the receiver and the other to the radiation. On the other hand, if a receiver having its natural frequency far away from the expected periodicities of the waves is employed, it is found too insensitive.[5] Worse yet, if the wave-train pursuing the path L_1ABM in Fig. 3 is a short heavily-damped one, while the natural oscillations of the receiver are of comparatively low frequency and slight decrement, the data will suggest that the wave-train is but slightly damped and has the frequency of the receiver.[6]

[5] Thermal receivers having natural frequencies far below those of the incident beams have been employed in studying wave-trains of much greater wavelengths and much more intense than these.

[6] This can be seen by considering an extreme case. Imagine that the wave is a single infinitely thin pulse, while the natural oscillations of the receiver are quite undamped. The pulse will be divided by the Boltzmann mirrors, so that two pulses

Clear smooth sine-like curves with the periodicity of the wave-train are obtained by using a receiver of which the natural frequency agrees with the fundamental frequency (or its octave) of the oscillator. Such curves are seen in Fig. 7; the two fundamental frequencies were



Set of curves, λ 4.2 to λ 27.

Fig. 7—Curves Obtained with a Receiver in Tune with the Oscillator (Topmost Curve with Receiver Tuned One Octave Below the Oscillator). (*Physical Review*)

will strike the receiver at a time-interval T; there will be nothing of the nature of interference. But if T happens to be an even-integer multiple of the half-period of the receiver, the second pulse will reinforce the oscillations started by the first, if it is an odd-integer multiple of the half-period, the second pulse will annul the vibration started by the first. Thus as the Boltzmann mirror is moved along, the receiver-reading will pass through maxima and minima with a spacing imposed by the characteristics of the receiver. In actual experiment this might happen if the frequency and the damping of the wave-train were much higher than those of the natural vibration of the receiver. On the other hand it does not appear that a frequency much higher than that of the wave-train could be simulated by any effect due to the receiver—an important point, in view of what follows.

in close agreement for all except the topmost of the curves, for which the fundamental of the oscillator corresponded to wavelength 4.2 mm. and that of the receiver to wavelength 8.4 mm.

The lowest wavelength mentioned by Nichols and Tear as having been manifested and visualized in this lucid fashion is 4.2 mm.; while, replacing the two mirrors of the Boltzmann interferometer by a set of eight mirrors forming an evenly-rising staircase or echelon, they obtained curves which in one instance indicated a fundamental of

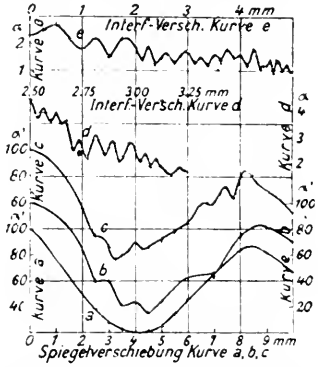
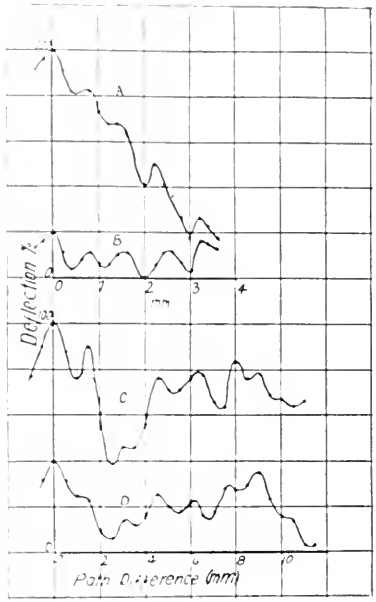


Fig. 8—Serrated Curves Indicating Very Short-Waved Components of the Wave Train. (*Annalen der Physik*)

1.8 mm. It would be a conservative, perhaps a too conservative, policy to regard this as the present limit of the spectrum of artificial electromagnetic waves.

Whether we may believe that rays lying beyond this limit have actually been generated depends upon the interpretation of certain narrow sharp serrations observed upon curves of the more uneven sort; for example, those of Fig. 8 (Möbius) and curves *A* and *C* of Fig. 9 (Nichols and Tear). If these are reliable indices of waves of corresponding wavelength in the mixed radiation from the oscillator, the frequencies in question must be considerably higher than the fundamental frequencies of the oscillators heretofore made; wavelengths ranging down to 0.1 mm., corresponding to frequencies ranging up to $3 \cdot 10^{12}$, have been inferred from such curves. If these are overtones emitted by the oscillator along with its fundamental, there would be little objection to extending the spectrum to cover them (although

it would be equivalent to considering a tenor's range as extending to the highest overtone which could be detected in any of his notes, which would certainly lead to astonishing results). Or they may be radiated by oscillations within particles of metal torn from the electrodes by the violence of the discharge—an idea suggested because



Harmonics.

Fig. 9.—Serrated Curves A and C Indicating Very Short-Waved Components of the Wave Train. ("Physical Review")

they are prone to appear after the spark-gap has been widened and the electrode-surfaces corroded by a succession of sparks. Or they might result from an excitation of molecules or atoms, in which case they should be regarded as belonging to the spectrum of natural rays, in the sense of my previous distinction.[7] Even so, if the serra-

[7] It is interesting to note that the opposite idea was put forward by Rubens (*l.c.*), that the rays of wavelengths 0.1 mm. and thereabouts emitted by the mercury arc might be due to oscillations in droplets of liquid mercury.

tions are truly due to waves issuing from the doublet and not to some unhappy peculiarity of the receiver—and the former alternative is considered the more probable one—then there is good reason for believing that the spectrum of artificial waves has been prolonged to overlap the spectrum of natural waves, and the lacuna is closed.[8]

The Discovery of Isotopes

Thirteen additional elements having been analyzed into isotopes by Aston, the moment is opportune for restating the two great series of discoveries which have disclosed the hidden law and the underlying unity of the chemical elements. Twenty-five years ago, the labors of chemists had resulted in setting apart about seventy-five distinct, unchangeable, non-interconvertible substances as "the elements"; and the ancient ambition to describe all forms of matter as combinations or modifications of a single, truly fundamental element must have seemed to be definitely frustrated. It is true that there were undeniable signs of a family relationship among the elements. They could be classified into groups of elements more or less alike in their properties; and when they were arranged in the order of their combining weights, there was distinctly a periodic variation of chemical

[8] Dr. Ernest Fox Nichols died suddenly on the twenty-ninth of April, 1924. A few days earlier he had very graciously offered to inform me of his latest work in extending the spectrum of artificial waves, hitherto unpublished except in brief reports before the Physical Society. He discussed the matter with his collaborator Dr. Tear, and to present his final formulation of his great achievement I can do no better than to quote verbatim a letter which Dr. Tear kindly wrote to me on April 29th:

"The most satisfactory data we have at present has been obtained with receivers whose fundamental wavelengths are long compared with those to be measured. The electrodes of our smallest oscillators are 0.1 mm. in diameter and 0.1 mm. long. The glass seal covers approximately one-half their length. The fundamental wavelength of such an oscillator is of the order of 1 mm. The distribution of the dielectric and the means of excitation are such however as to accent certain harmonics and to suppress the fundamental and other frequencies. The interference curves show then the presence of one high frequency, usually the second or fourth harmonic, plus the low frequency of the receiver. The interference persists for three or four cycles and is reproducible, although the construction of such minute seals introduces the element of chance, frequently making it necessary to construct several oscillators before finding one having the right proportions of bare and glass-covered electrode-surface to bring out one frequency and suppress the remainder.

"It is in a way perplexing that although chance proportions of glass and metal bring out one harmonic to a greater or less degree, the fundamentals of these smallest oscillators do not show up at all. It is of interest, too, to note that a sheet of glass 0.2 mm. thick, such as the seals are made of, transmits but 25% of the 0.32 mm.-radiation from the mercury arc. We have been led to the interpretation that the particular standing waves which can exist upon these small oscillators are determined by the location of the glass-oil boundary-surface, and that the predominant wavelength is the fundamental wavelength of that part of the oscillator which is in oil between the two glass-oil surfaces. *The wavelengths which we have isolated in this way extend to the 0.22 mm. limit* which we reported at Boston." (That is, at the Boston meeting of the Physical Society, December, 1922. Italics mine.—K. K. D.)

and physical properties in passing along the line. Indeed the periodicity was so clear that in three instances when the order of two consecutive elements was such as to damage the periodic law, chemists simply reversed the order—putting argon before potassium, cobalt before nickel, tellurium before iodine, thus testifying to a faith that there must be something governing the nature of the chemical elements more fundamental than combining weights. Furthermore, in several instances where the periodic law implied that there ought to be an additional element between two apparently consecutive ones, chemists left a vacant space between the two for an element presumed to be existent but unknown; and some of these elements were subsequently discovered, thus justifying the faith in the most impressive way. But of the nature of this fundamental something, there was no inkling.

It had been suggested at one time that all atoms are built up of hydrogen atoms. But the most accurate measurements placed it beyond doubt that the chemical combining weights of the elements are not, in every case, integer multiples of the combining weight of hydrogen, nor of any other common divisor large enough to have a physical meaning. As it was universally assumed that the weight of the ultimate particles of an element is equal to its combining weight multiplied by some universal factor, this fact seemed to disprove the suggestion. Yet on the other hand the measurements established a rule that the combining weight of many of the elements—far too many to be explained as due merely to chance—are integer multiples of a common unit which is $\frac{1}{16}$ of the combining weight of oxygen. This can be illustrated from any group of elements, for example from the first ten of the periodic table:

H	He	Li	Be	B	C	N	O	F	Ne
1.008	4.00	6.94	9.02	10.83	12.00	14.01	16.00	19.0	20.20

out of which group of numbers eight are integer multiples of the unit 1.00, within observational error; while four—the combining weights of hydrogen, lithium, boron and neon—certainly are not. We are confronted with a manifest rule restricted by undeniable exceptions—the most stimulating situation which can arise in a science.

Suddenly the exceptions to the rule were all explained away, and the mystery vanished with a completeness which we hope that some of the other mysteries of physics will some day emulate. The trouble was simply that everyone has assumed, with an indifference to the other alternative which now seems strange,[1] that all the atoms of an

[1] Compare Aston's historical review (*Isotopes*, pp. 1–6). Crookes very definitely suggested a multiplicity of atomic weights in 1886.

element have the same weight, which (multiplied by the proper universal factor) is the combining weight of the element; whereas now it is known that some of the elements have two or several different kinds of atoms apiece, with different weights, of which the observed combining weight of the element is merely an average. The combining weight of an element as observed in ordinary chemical experiments has no general right to the title of *atomic weight*; only in special instances may the two be identified. The elements of which the combining weights are integers—meaning, integer multiples of $\frac{1}{16}$ of the combining weight of oxygen—consist of atoms of a single kind, the weight of which is truly and accurately given by the combining weight of the substance. The others, or those of them which have been analyzed, are mixtures of atoms of different kinds, the weight of no one of which is given by the combining weight of the element. Wherever it is actually the mass of an atom which is measured by the chemical method, the rule is verified; where the rule is apparently infringed, the quantity measured is merely a misleading average, and not the mass of an atom at all. When, therefore, the rule is restated to apply only to those combining weights which are truly atomic weights, the conspicuous exceptions no longer militate against it, and the supposition that all atoms may be built of hydrogen atoms is strongly reinforced.

When J. J. Thomson developed the technique of his "positive-ray analysis" by which he measured the masses of fast-flying charged atoms and molecules, he was unknowingly preparing the way for ascertaining how many different kinds of atoms belong to a single element. In these classical experiments the ionized particles were those existing in a rarefied gas traversed by an electrical discharge, and drawn to the cathode by the strong field maintaining the discharge; through a narrow perforation in the cathode, a thin pencil of the ions passed into a chamber where it was subject to crossed electric and magnetic fields. These fields resolved it into a number of separated and separately-directed pencils, each containing exclusively atoms (or molecules) of a single uniform mass, which could be deduced from the location of the trace made by the pencil upon a photographic plate.[10] The method was designed by Thomson as a sensitive, indeed a supersensitive, method of chemical analysis, by

[10] The actuality is somewhat more complex, as a distinct pencil is obtained for each value of the charge-mass ratio E/m, and it is this ratio which is deducible from the location of the pencil. However E is either the electron-charge e or a small integer multiple of it (occasionally, but rarely, as great as $8e$), and the multiplicity of pencils corresponding to different values of E and a single value of m seems to be an actual advantage to the experienced interpreter of such data.

which gases present even in small proportions in a discharge-tube could be detected and identified. The first trials were naturally made upon discharge-tubes containing the commoner gases, which as it happens nearly all consist of one kind of atom (or molecule) apiece: oxygen, nitrogen, hydrogen, carbon dioxide, carbon monoxide. This retarded the great discovery. But when neon, a gas of presumed atomic weight 20.2, was introduced into the tube in 1912, Thomson observed two pencils, of atoms of masses about 20 and 22, respectively, where he had expected to see but one consisting of atoms of mass about 20.2.[11]

This observation was not immediately interpreted as we now interpret it. The mysterious pencil might have consisted of molecules of CO_2 of mass 44 bearing a double charge, or of molecules of a hitherto unknown compound NeH_2. These possibilities were tested by appropriate experiments and discarded, and then for a time the gas of atomic mass 22 was apparently regarded as a new element distinct from neon and fortuitously mixed with it.

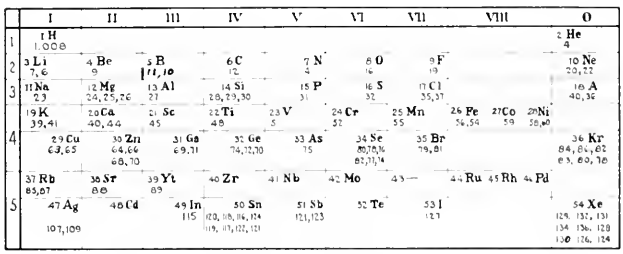
F. W. Aston undertook the attempt to separate the two gases, but they were so entirely alike in their properties that no success whatever was attained by fractional distillation and little by diffusion. This was Aston's entry into this field, and in a celebrated series of researches, soon interrupted by the war but resumed after six years and still continuing, he associated his name forever with the analysis of elements into the different kinds of atoms of which they consist.

Of the improvements which Aston made in the method of measuring the masses of charged particles, as of the details of Thomson's original method and of Dempster's method, it is hardly necessary to speak; for they have been admirably described, with reproductions of photographs, in several recent books.[12] The problem of generating ions of the elements to be analyzed became progressively harder to solve. The elements gaseous at room-temperature were easily investigated, and those of which a high vapor density could be produced either of the element or of one of its compounds, without overheating the tube, were also tractable; but when these elements had all been tested the resistance to further advance became formidable. Ions of the thirteen elements lately analyzed were formed as *anode rays*: that is, they were charged atoms expelled from the anode of a discharge-tube

[11] Neon by virtue of its well-known chemical inertness has no "combining" weight, but its average molecular weight was determined from its density by Watson, using Avogadro's principle, as 20.200. Thomson's earliest experiments were not delicate enough to distinguish whether the atoms in the former of the two pencils were of mass 20.0 or of mass 20.2, but the difference between either and 22 was unmistakable.

[12] Notably in Aston's own book *Isotopes* and in Andrade's *The Structure of the Atom*.

during the discharge—not ionized atoms of a gas sustaining the discharge, as previously—and drawn to and through the cathode by the entire voltage across the tube. The anode of the tube must be made in a special manner; in Aston's experiments it consists of a "paste" made of graphite, of lithium iodide, of a halogen salt of the metal to be analyzed, and sometimes of other salts as well. Ions of the other elements in the paste and from the gas in the discharge mingle with the desired ions in the pencil which shoots through the cathode perforation, but this is no inconvenience, quite the reverse, as the traces



Additional elements 55 Cs, one isotope at 133
80 Hg, isotopes at 202, 204, and in the range 197-200

Fig. 10—The First Six Rows of the Periodic Table of the Elements, Showing the Atomic Masses of the Isotopes of the Elements which Have Been Analyzed. The Data Come from Aston's Tabulations

which they leave on the plate are convenient *points de repère* for fixing the exact location of the traces left by the ions being analyzed. In this manner the elements scandium, titanium, vanadium, chromium, manganese, cobalt, copper, gallium, germanium, strontium, yttrium, silver and indium were studied—thirteen altogether, bringing up to forty-seven the number of elements which have been analyzed in the fourteen years since neon was first discovered to be multiple.

All of these forty-seven elements except two lie in the first five periods of the periodic table, and they have been written out in the tabular form in Fig. 10. The symbol of each element is preceded by its atomic number, and below the symbol lies not the combining weight of the element as in the tables one usually sees hanging on the walls of chemical lecture-rooms, but the ensemble of the atomic weights of its various kinds of atoms—the atomic masses of its *isotopes*, as the term is. Where no number or set of numbers is given, the analysis has not yet been made. Of the first fifty-five elements of

the table, all have been analyzed except nine; but of the next twenty-seven elements, only one (mercury) has been analyzed. These heavier elements and their compounds seem generally to be non-volatile and so impregnable by the original method; while they are difficult, if not impossible, to examine by Aston's new scheme, as the traces of the ions upon the plate become fainter with increasing mass, and are already extremely faint for the elements in the fifth row of the table.

Among the eight known elements beyond the eighty-second, every one has atoms of several different kinds, alike in physical and chemical properties but different, it is presumed, in mass; but they differ also in another quality, a much more striking quality—they differ in their degree of instability. Out of a great number of atoms of a radioactive substance, existing at a moment t, one-half will have disintegrated at a subsequent instant $t+T$; the interval T, which is called the *half-period* of the substance, is the measure of its instability. Like the atomic mass, this half-period may vary from one kind of atom to another, though both kinds have almost identical chemical and physical properties and belong to the same element. The three isotopes of the eighty-sixth element, "emanation," have three entirely distinct half-periods: 54 seconds, 3.85 days and 11.2 days. Moreover, not only the rate but the manner of disintegration may be different for different isotopes of a single element. The six kinds of atoms which share the ninetieth place in the periodic table display this diversity of properties:

Uranium X_1 has a half-period of 23.8 days and its atoms emit electrons and electromagnetic waves when breaking up;

Uranium Y has a half-period of 24.6 hours and emits electrons;

Ionium has a half-period of 9×10^4 years and emits helium nuclei;

Thorium has a half-period of 2.2×10^{10} years and emits helium nuclei;

Radiothorium has a half-period of 1.90 years and emits helium nuclei;

Radioactinium has a half-period of 19 days and emits helium nuclei.

Nor must it be supposed that if each of two isotopes is stated to emit helium nuclei, they are in that respect identical; for the energies of the emitted nuclei generally vary from one isotope to another, so that every one of the six kinds of atoms listed above differs from every other not only in the rate but also in the manner of its disintegration and likewise in its ancestry and its posterity, in the

genealogical line to which it belongs. These are the conspicuous differences between the isotopes comprised in a single element, when the element is one of the eight final ones of the periodic table; and they are so conspicuous and impressive that the individual isotopes

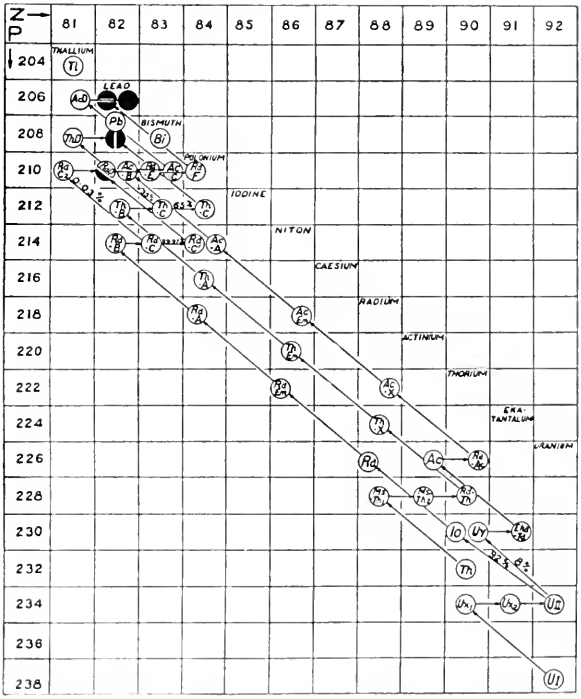


Fig. 11 Table of the Elements from the Eighty-Second to the Ninety-Second (Inclusive) Showing the Names and the Presumed Atomic Masses of Their Isotopes. (Andrade, *The Structure of the Atom*)

enjoy individual names, while the element to which they all belong usually has no all-embracing name of its own, quite the opposite of the state of affairs among the stable elements, each of which has its own name while the isotopes composing it are known only by the numbers giving their atomic masses. Fig. 11 displays these last

eight elements and the preceding one (lead) each in a column of its own marked with its atomic number (and for identification the name of some element in the same column of the periodic table), while the mass of each kind of atom is given by its elevation above the bottom of the figure (the values are written along the vertical axis).[13]

In these tabulations of Fig. 10 and Fig. 11, all the numbers representing atomic masses are written as integers. The conspicuous post-decimal figures occurring in the sequence of combining weights are absent; the notorious 35.45 of chlorine, the 24.32 of magnesium, the 10.83 of boron have vanished from the scene. Are then the masses of all atoms really integer multiples of $\frac{1}{16}$ of the mass of the oxygen atom, using "really" in its only significant sense of "within the uncertainty of observation?" Or do some of them deviate appreciably from the rule? The trial can be made most exactly upon the lightest elements, as for these a given deviation from an integer value would bulk as a larger percentage of the total mass, which is the measured quantity, than it would for the heavy elements. It is performed by mingling the ions under test with ions of oxygen, or of some other element, preferably one which has previously been compared with oxygen; the locations of the traces of the two pencils of ions upon the photographic plate are compared. Mingling lithium ions with carbon ions, Aston finds that the masses of the two kinds of lithium atoms stand to the mass of the carbon atom as

$$(7.006 \pm .005) : 12.000 \text{ and } (6.008 \pm .005) : 12.000$$

and if the mass of the carbon atom is exactly $\frac{12}{16}$ that of the oxygen atom, then the masses of the lithium atoms are very slightly distinct from $\frac{7}{16}$ and $\frac{6}{16}$ of the oxygen mass (for, little as the difference exceeds the uncertainty of experiment, Aston regards it as real). Beryllium, however, yielded the values 9.003 and 9.001 —indistinguishable experimentally from 9.000— in two separate experiments, in terms of the same assumed value 12.000 for carbon. Farther along in the

[13] The atomic masses of these different kinds of atoms are largely hypothetical. They have been measured for four single isotopes belonging to four distinct elements: radium (number 88, mass 226), radium emanation or niton (number 86, mass 222), thorium (number 90, mass 232), and uranium (number 92, mass 238). Measurements have also been made upon samples of lead believed to be composed almost entirely of a single isotope, giving 206 for one kind of atom and 208 for another. Each of the other atoms is a descendant of one or two of the four first named atoms, and its atomic mass is calculated by subtracting, from the atomic mass of its ancestor, the masses of all the fragments which dropped away from the earlier atom during its evolution. This procedure is confirmed by comparing the measured values for uranium, radium, radium emanation, and one sample of lead, which all belong to the same line of evolution; and the measured values for thorium and for another sample of lead which is descended from thorium.

procession of elements, comparisons with the oxygen atom become difficult; but adjacent elements can be intercompared. The eight sorts of tin atoms lie next to the nine sorts of xenon atoms, the most massive kind of tin agreeing closely in weight with the least massive kind of xenon. When atoms of gaseous xenon and molecules of a volatile compound of tin are mixed together in the discharge-tube, the beam of ions issuing through the cathode-perforation is resolved into seventeen pencils; and the seventeen traces upon the plate are so placed that the masses of the seventeen atoms cannot all be integer multiples of a common unit of the order of $\frac{1}{16}$ the oxygen mass.[14] Either the tin atoms or the xenon atoms deviate appreciably from the rule, or possibly both do.

So the common history of great sweeping discoveries in science seems to repeat itself; the simplicity of the principle first announced is gradually marred, its sharp lines become a trifle hazy and vague, as experiments are multiplied and refined. Yet the principle does not for that lose its character or its importance; the deviations of the new group of values from integer numbers are small compared to those of the old one, and promise to amplify the physical meaning of the rule instead of restricting it. We should be less prepared to accept them, were there not one of them at the very root of the system of elements; for the mass of the hydrogen atom is not exactly the $\frac{1}{16}$ of the oxygen mass which was taken for the fundamental unit mass of the system of atoms, but is 1.008 16 of it. This seems embarrassing; the bricks of which we intended to say that the atomic structures are built turn out to be smaller than the sample brick. But the embarrassment can be removed; for it can be shown that of the mass of the hydrogen atom is altogether electromagnetic, then the total mass of a group of such atoms crowded closely together must be inferior to the sum of the masses of the individual atoms when far apart. Therefore, small deviations from the rule of integer masses are to be anticipated, and may be expected to serve as a most valuable contrôle of proposed models of atom-nuclei, when the epoch of quantitative spatial models arrives. This epoch may be distant; or we may be upon the verge of it.

We have admitted, then, that the combining weight of an element, being in general not its atomic mass but the average of the masses of several kinds of atoms, and a *weighted* average at that, does not have

[14] The experiment was performed with a tube containing the gaseous compound tin tetramethide $(SnCH_3)_4$ and some xenon from a previous experiment. Eight pencils of $SnCH_3$ ions were observed, consisting of molecules comprising tin atoms of the eight different kinds; molecules containing tin atoms of mass 120 would have a total mass of 135, and hence a pencil containing them would have fallen just midway between the pencils of xenon atoms of masses 134 and 136, respectively; actually it fell distinctly off-centre.

the profound physical significance it once seemed to possess. But this is not all; we must further concede that even the mass of the atom or the ensemble of masses of the atoms of an element is not by any means so distinctive and important a quality of the element as one would expect. Not only may one element have atoms of several different masses, but two distinct elements may have atoms of, so far as we can distinguish, the same mass; argon and calcium, selenium and krypton, tin and xenon. Now if an atom of the gaseous and inert argon may have the same mass as an atom of the metallic and active calcium, we cannot evade the conclusion that the mass of an atom is, in the terms of logic, an *accidental* property of the atom rather than an essential one. There must be some fundamental and essential feature or quality of the atom, which determines its ensemble of physical and chemical properties, and which is not the atomic mass; perhaps this quality determines the atomic mass as well, but certainly not in so rigorous a manner that one value of atomic mass corresponds invariably to one set of chemical and physical properties, and vice versa. This fundamental feature of the atom we recognize as the charge upon its nucleus, which, expressed as a multiple of the electron-charge e (of which it must be an integer multiple[15]) is also the number of electrons accompanying the nucleus, and the *atomic number* of the element.

This nuclear charge, or (cardinal) electron-number, or (ordinal) atomic number, is the same for all the atoms of a single element, and never the same for two atoms of different elements. It is 50 for all of the eight kinds of atoms of tin, and 54 for all of the nine kinds of atoms of xenon. It is 18 for all atoms of argon and 20 for all atoms of calcium, though some atoms of the one have the same weight (within one part in a thousand, Aston says) as some atoms of the other. It is 26 for all atoms of cobalt and 27 for all atoms of nickel, though most of the atoms of nickel are lighter and a few heavier than the atoms of cobalt. It is the true basis for the ordering of the elements, of which the ordering of the atomic masses is but an imperfect and distorted (though not a badly distorted) imitation.

Five observations or assemblages of observations, made in fields of physics separated almost as widely as any five fields could be, sustain this principle; and, combined with its philosophical attractiveness for the idea of arranging the elements in a single procession and attaching consecutive integer numbers to their fundamental qualities is as irresistibly attractive as a scientific idea can be make it about

[15] Otherwise the nuclear charge could not be exactly balanced by the charges of the environing electrons

as certain as any principle not dealing with things which can be seen and handled. I shall mention them briefly, in a nearly chronological order.

The direct measurement of the charge of the helium nucleus. Rutherford and Regener independently measured the charge on the alpha-particle in the simplest, most direct and most incontrovertible way; they counted the alpha-particles emitted from a sample of a radioactive element in a given time, and measured the total charge they carried away from it, and divided the one datum by the other. Rutherford obtained twice $4.65 \cdot 10^{-10}$ (electrostatic units) for the charge of the individual particles; Regener obtained twice $4.79 \cdot 10^{-10}$. The agreement of the latter value with twice Millikan's standard value of the electron-charge ($4.774 \cdot 10^{-10}$) is magical; the agreement of the former value is also good. Though this is an average value for a great number of particles, the fact that a beam of alpha-particles is not spread or split by a magnetic field proves that each has the same charge (at least, to be perfectly precise, the same charge-to-mass ratio). It is established that the alpha-particle is the bare helium nucleus.[16]

The determination of nuclear charges by the scattering of alpha-particles. When a beam of alpha-particles is played against a sheet of metal foil, the nuclei of the metal atoms deflect the alpha-particles passing very close to them, by virtue of the electrostatic repulsion between the charge $+2e$ on the alpha-particle and the charge $+Ne$ on the nucleus of the metal atom. The distribution-in-angle of the scattered alpha-particles can be calculated, assuming that the action of the metal nuclei is not complicated by any forces due to the electrons surrounding them. The distribution-in-angle actually observed agrees in form with the calculated one; this proves not that there are no electrons surrounding the metal nuclei, but that there is a vacant space around each nucleus, wide enough so that the major part of the deflection of an alpha-particle takes place within it. All this was discussed in the second article of this series. The form of the distribution-in-angle and its variation with the speed of the alpha-particles prove the existence of the atom-nuclei, of their positive

[16] For helium gas, evinced by its spectrum, appears in a tube into which alpha-particles are fired through the wall, and is exuded from a piece of metal which is melted after alpha-particles have been shot into it (Rutherford's experiments); furthermore impacts between alpha-particles and atoms of helium gas show them to be of the same mass (Blackett's experiments), and the value of the e/m ratio is correct for doubly-ionized helium atoms but not for any other admissible variety of atom. And the radius of the alpha-particles, calculated from the experiments on scattering, is smaller by several orders of magnitude, than the effective radius of any known atom having electrons in addition to its nucleus.

charges, of the vacant space around them; and if the percentage of scattered particles is measured absolutely, the absolute value of the charge of the nucleus can be calculated. These values have actually been obtained:[17]

Platinum: nuclear charge	$(77.1 \pm 1)e$	(Chadwick)
Silver:	$(46.3 \pm 0.7)e$	(Chadwick)
Copper:	$(29.3 \pm 0.5)e$	(Chadwick)
Argon:	$19e$	(Auger and Perrin)
"Air"	$6.5e$	(C. T. R. Wilson)

Bohr's interpretation of the spectra of hydrogen and ionized helium. There is a complete and perfect agreement between the observed frequencies in the spectra of hydrogen and ionized helium, and the frequencies predicted by Bohr. An essential feature of Bohr's theory is that the charge on the nucleus of the hydrogen atom is assumed to be e, and the charge on the nucleus of the helium atom to be $2e$. As there is no other element of which the spectrum has been perfectly and completely explained by Bohr's theory (or any other) this affirmation cannot be extended beyond hydrogen and helium.

Thus we have excellent evidence from three distinct sources that the nuclear charge of helium, the second element of the periodic table, is $2e$; excellent evidence from two sources that the nuclear charge of the first element, hydrogen, is e; and good evidence by the alpha-ray method that the nuclear charges of the 18th, 29th, 46th and 78th element are as close to $18e$, $29e$, $46e$, and $78e$ as to any other integer multiple of e. In addition, there is evidence from two more sources that, in passing from one element to the next along the procession of elements, one finds the nuclear charge augmented by the amount e at each step; thus completing the itemized evidences foregoing by a process somewhat like what is called "mathematical induction."

The displacement-law of Fajans and Soddy. When an atom-nucleus of a radioactive element disintegrates by shooting off an electron bearing a charge $-e$, the residuum is found to be a nucleus of an element one step farther up in the procession of elements. When an atom-nucleus disintegrates by shooting off an alpha-particle bearing a

[17] The earliest experiments (discussed by Rutherford in 1911) demonstrated that for several metals the nuclear charge (measured in terms of e) was about one-half the atomic weight, and those of Geiger and Marsden 1913 were arranged primarily to demonstrate the validity of the concept of the nuclear atom, but confirmed that statement for gold. Chadwick repeated these experiments upon *Pt*, *Ag* and *Cu* with the object of determining the nuclear charge as accurately as possible. The values for argon and "air" were determined by what is in principle the same method though in a very different form; the former with alpha-particles, the latter with fast electrons.

charge $+2e$, the residuum is found to be a nucleus of an element two steps farther down in the procession of elements. Thus in passing from one element to the next above it, the nuclear charge is found to be augmented by e. This law is deduced from numerous observations on the elements beyond the eighty-first.

Moseley's law. The square root of the frequency of the $K\alpha$-ray (a prominent and easily-identified member of the X-ray spectrum) increases by a constant amount in passing from one element to the next above it. This law is valid from the twelfth to the ninety-second element in the periodic table. The same law governs, though not with such entire accuracy, the other identifiable members of the X-ray spectrum.

Apart from all interpretation, Moseley's law means that there is a certain important measurable quantity which is very characteristic of the elements and increases uniformly and steadily from one to the next, over almost the entire procession. The mere existence of such a quantity inspires confidence that there is a true physical seriation of the elements, but by interpretation a great deal more can be added. Bohr's theory of the atoms of hydrogen and ionized helium lead to this result: when a single electron forms an atomic system with a nucleus of charge Ne, one of the frequencies which this system can radiate—and the frequency which, on the whole, it would oftenest and most intensely radiate—is equal to

$$\nu = \frac{3}{4} RN^2, \qquad R = 2\pi^2 m\, e^4 / h^3. \tag{1}$$

This is verified for hydrogen and ionized helium, each of these atoms consisting of a nucleus and a single electron. No other such atom has yet been isolated and made to radiate. But we might imagine that in a massive atom containing many electrons, one lies deep down beneath the others, and revolves by itself in the field of the nucleus, undisturbed by the rest. In this case there would be an X-ray frequency emitted by the atom, given by (1). The difference between the values of the square root of this frequency for consecutive elements would be constant and equal to

$$\Delta = \sqrt{\frac{3}{4} R}. \tag{2}$$

Now the observed constant difference between the values of the square root of the $K\alpha$ frequency for consecutive elements does conform to (2). But the actual value of the frequency does not conform to (1)

unless we get the quantity N equal, not to the order-number of the element in the procession, but to the order-number minus one.

Does this mean that the nuclear charge of the nth element in the periodic table is $(n-1)e$ for all the values of n exceeding 11 (the values for which Moseley's law holds)? I fear this could not be contradicted from the direct experimental evidence, for Chadwick's values of the nuclear charges for the elements $n-78$, 47, 29 fall just short of being exact enough to prove that they are $78e$, $47e$, $29e$ instead of $77e$, $46e$, $28e$, respectively. However, we should do too much violence to the beauty of the principle if we admitted that there are only eight values between $2e$ and $11e$ to be distributed among the nuclear charges of the nine elements between helium and magnesium, and happily it is not necessary, for the apparent discordance can plausibly be blamed upon too simpe a view of the internal economy of the atom which we took in deriving equation (1). Instead of assuming that the deepest-lying electron of the atom revolves in an otherwise vacant space surrounding the nucleus, wide enough to contain the first two of its permissible orbits, we should do better to assume that there are several deep-lying electrons similarly placed and interacting with one another, or at least that there is no single deepest-lying electron too far inward to be affected by the others. The effect of thus changing the assumption is to change the calculated value of the $K\alpha$-frequency, for an atom of nuclear change Ne, from the value (1) to a value $\frac{3}{4}R(N-k)^2$; in which k depends on the particular configuration assumed for the internal electrons. It is clear, therefore, that we are in no wise compelled by Moseley's law to conclude that the nuclear change of the atom of the nth element is $(n-1)e$ when $n>11$, and may continue to accept the much more satisfying principle that the nuclear change of the nth element is ne.[15]

Before stating the conclusion let me restate the evidence in a briefer form and an altered order. Originally the elements were arranged in the order of their combining weights. It was seen that when they are arranged in this way, there is a periodic variation of the ensemble of chemical and physical properties from element to element. But to make the periodic variation quite smooth and unbroken, it was found necessary to violate the order of the combining weights at several places in the series; three pairs of con-

[15] The agreement with experiment indeed becomes very good, at least over a certain range of elements, if we assume that there are normally 3 electrons in the innermost or one-quantum ring and nine in the second or two-quantum ring orbit (J. Kroo). But nobody wants to accept this particular repartition of electrons, and it is customary to assume that the inner orbits are mostly elliptical. But it would be gratifying to attain a quantitatively successful theory.

secutive elements had to be reversed, and at several points it was necessary to leave vacant spaces between apparently-adjacent elements, imagining undiscovered ones to separate them. Thus it became clear that the true arrangement of the elements was controlled by something deeper and more fundamental than the combining weights; yet there was no adequate reason for preferring one of the measurable physical or chemical properties above all the others as the fundamental one. Moseley then discovered that the square root of the most conspicuous X-ray frequency increased at a steady and even pace from one element to the next, throughout almost the entire list of elements. Where the order of combining weights disagreed with the order of physical and chemical properties, the order of X-ray frequencies agreed with the latter and not with the former; where the succession of chemical and physical properties suggested that an element was missing from the list, the excessive leap of the root of the X-ray frequency in passing from the element below to the element above the suspected gap gave a striking confirmation. This important quality of the elements, advancing by equal steps from one to the next, testified far more impressively than the periodic variations of the various chemical and physical qualities to the close affiliation among them.

Measurements of the deflections of alpha-particles by atoms had shown that the atom has a massive nucleus bearing a positive charge; as there are also electrons surrounding the nucleus, and as no one has proved the existence of negative electricity otherwise than in electrons, it was inevitable to believe that the positive nuclear charge is balanced by and balances the charges of the surrounding electrons, and so is an integer multiple of the electron-charge e. Moseley's law could be interpreted to mean that the nuclear charge increases by e in passing from one element to the next. Fajans and Soddy had already found that when one of the radioactive elements is transformed into another, the transformation is always such that an increase of e in nuclear charge goes with an advance of one step along the series of elements. Therefore, it would be possible to assign the nuclear charges of all the elements if the nuclear charge of one, or preferably of several, could be absolutely determined. The experiments upon scattered alpha-particles did show for several elements that the nuclear charge of the nth element is at least as close to ne as to any other integer multiple of e; direct measurement of the nuclear charge of the second element showed that it is quite accurately $2e$; and Bohr's theory, of which the interpretation of Moseley's law was an offshoot, derived its own successes partly from the essential as-

sumption that the nuclear charges of the first and second element are e and $2e$, respectively.

Meanwhile the combining weights, without losing their practical utility, were slipping out of the prominence into which they had been forced. It was discovered that they were not always to be identified with the atomic weights; that an element might have several kinds of atoms; that even the masses of these atoms were not absolute characteristics of the elements, as two very different elements might have atoms of apparently identical mass. In Remy de Gourmont's phrase, there occurred a *dissociation of ideas*; the idea of atomic weight was dis-associated from the idea of element, and the idea of atomic number supplanted it. The eighty-seven (now the eighty-eight) known elements formed themselves into a procession, which is a procession of atoms bearing eighty-eight of the ninety-two admissible nuclear charges between e and $92e$, and possessing consecutively all except four of the possible electron-families ranging in number from one to ninety-two. That at least eighty-eight out of these ninety-two conceivable atoms should actually exist and have been discovered, may seem strange; one might perhaps have expected that a stable nucleus with a net charge of ne could be built only for an occasional value of n; but among the first eighty-two integers there are certainly not more than two, perhaps none, which are not represented by durable nuclei; and among the next ten at least eight are represented by not-too-transient nuclei. We have also seen that nuclei with certain values of charge, $54e$ or $90e$ for example, can be constructed in several different ways. These problems of nuclear structure are, however, problems for the future. What does seem established at the present moment is, that if we could determine the properties of the system formed by n electrons and a nucleus of charge ne, we should know all the properties of the elements except a very few having to do with intra-nuclear events. As the only case thus far successfully dealt with is the case $n = 1$, and we cannot even explain what happens when two such atoms combine, this is not meant as an augury of an early complete liquidation of the mysteries of physics. Nevertheless, we have good reason to believe that, though ours is doubtless not the generation which will complete the solution of the problem of the atom, it is the first to which the nature of the problem has been revealed.

REFERENCES

F. W. Aston: Phil. Mag. *47*, pp. 385–400 (1924); *45*, pp. 934–945 (1923). *Isotopes* (London, 1922).

P. Auger and F. Perrin: Comptes Rendus, *175*, pp. 349–343 (1922).

J. Chadwick: Phil. Mag. *40*, pp. 734–746 (1920).

H. Geiger and E. Marsden: Phil. Mag. *25*, pp. 604–623 (1913).

J. Kroo: Physikal. ZS. *19*, pp. 307–331 (1918).

F. Mobius: Ann. der Phys. *64*, pp. 293–322 (1920).

H. G. J. Moseley: Phil. Mag. *27*, pp. 703–713 (1914); *26*, pp. 1024–1034 (1913).

E. F. Nichols and J. D. Tear: Physical Review *21*, p. 378 and pp. 587–610 (1923).

E. Regener: Sitzungsber. Berlin Academy, 1909, pp. 948–965.

E. Rutherford: Phil. Mag. *21*, pp. 669–688 (1911) (evidence that nuclear charge is about $\frac{1}{2}$ atomic weight). Proc. Roy. Soc. *81A*, pp. 162–173 (1908) (evidence that nuclear charge of *He* is $2e$). Article "Radioactivity," Encyc. Brit. *32*, pp. 219–223 (1922) (isotopes among the radioactive elements).

C. T. R. Wilson: Proc. Roy. Soc. *104A*, pp. 192–212 (1923).

Some Very Long Telephone Circuits of the Bell System

By H. H. NANCE

RECENT papers[1] have discussed at length the use of toll cables for the handling of certain long distance traffic. These cables, which are being used in areas of dense traffic, have been made possible by many developments in cable design, repeaters and loading coils. Coincident with these developments have gone others which are finding their application in the extensive establishment of improved open wire circuits for use over very long distances. The purpose of the present paper is to discuss some of the considerations involved in the overall design and maintenance of these very long open wire circuits. These circuits are often referred to as "backbone" circuits and supply a network of trunk lines for the entire Bell System. The most important of these routes are shown in Fig. 1.

The first transcontinental line was completed in the summer of 1914 and early in the following year three transcontinental telephone circuits were placed in commercial service. These circuits were constructed of copper wire 165 mils in diameter loaded with 250 millihenry coils at intervals of 7.88 miles and had telephone repeaters located at points about 500 miles apart. The opening of these first circuits, while marking a most important stage in the progress of long distance telephony, has been followed by many developments which have made possible increased overall transmission efficiency and improved quality. A discussion of these developments is given in a paper on "Telephone Transmission Over Long Distances," by H. S. Osborne.[2]

Two outstanding characteristics of these new open wire circuits are that they are non-loaded and that the repeaters are of an improved type, the number being increased in consequence of the higher attenuation. With these long non-loaded circuits increased speed of propagation and smoother characteristics are obtained resulting in less echo effect and better volume. Better attenuation-frequency characteristics are obtained and the quality is further improved due to the elimination, to a large extent, of transients. Changes in line attenua-

[1] "Philadelphia-Pittsburgh Section of New York-Chicago Cable," by J. J. Pilliod, *Bell System Technical Journal*, Vol. I, No. 1, July, 1922; *Journal of A.I.E.E.*, August, 1922. "Telephone Transmission Over Long Cable Circuits," by A. B. Clark, *Journal A.I.E.E.*, January, 1923; *Bell System Technical Journal*, Vol. II, No. 1, January, 1923.

[2] For detail discussion see paper on "Telephone Transmission Over Long Distances," by H. S. Osborne, *Journal A.I.E.E.*, Vol. XLII, No. 10, October, 1923.

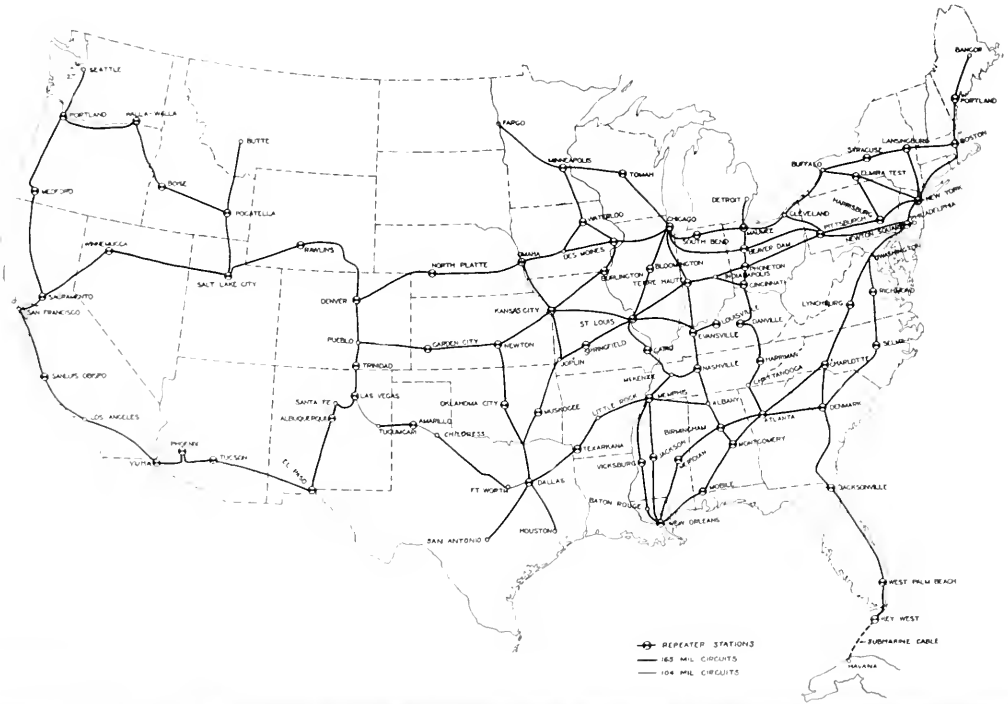

SEATTLE
PORTLAND
WALLA WALLA
BUTTE
BOISE
POCATELLO
SALT LAKE CITY
RAWLINS
DENVER
NORTH PLATTE
OMAHA
FARGO
MINNEAPOLIS
TOMAH
DES MOINES
WATERLOO
CHICAGO
SOUTH BEND
DETROIT
MILWAUKEE
BEAVER DAM
BLOOMINGTON
BURLINGTON
TERRE HAUTE
INDIANAPOLIS
CINCINNATI
KANSAS CITY
ST LOUIS
LOUISVILLE
DANVILLE
EVANSVILLE
NASHVILLE
HARRIMAN
CHATTANOOGA
CHARLOTTE
SELMA
MACON
ATLANTA
ALBANY
MEMPHIS
LITTLE ROCK
MERKEL
SPRINGFIELD
JOPLIN
PUEBLO
CANON CITY
NEWTON
TRINIDAD
LAS VEGAS
SANTA FE
ALBUQUERQUE
OKLAHOMA CITY
MUSKOGEE
AMARILLO
TUCUMCARI
CHILDRESS
TEXARKANA
VICKSBURG
JACKSON
MONTGOMERY
MOBILE
BIRMINGHAM
DALLAS
FT WORTH
BATON ROUGE
NEW ORLEANS
HOUSTON
SAN ANTONIO
EL PASO
PHOENIX
TUCSON
YUMA
SACRAMENTO
SAN FRANCISCO
SAN LUIS OBISPO
LOS ANGELES
BANGOR
PORTLAND
BOSTON
LANCASTER
SYRACUSE
BUFFALO
OLMSTED TEST
NEW YORK
PHILADELPHIA
HARRISBURG
PITTSBURGH
CLEVELAND
NEWTON SQUARE
WASHINGTON
RICHMOND
LYNCHBURG
JACKSONVILLE
WEST PALM BEACH
KEY WEST
SUBMARINE CABLE
HAVANA
REPEATER STATIONS
165 MIL CIRCUITS
104 MIL CIRCUITS

Fig. 1

tion with weather conditions are also considerably reduced. Furthermore, the use of this type of circuit fits in with the application of carrier current systems for which it is advantageous to use non-loaded 165 mil circuits where these are available.[3]

With the improved repeaters and balancing networks it is possible to obtain a higher degree of balance at the various repeater points. The improved transmission characteristics of these repeaters also contribute toward better quality. Both of these improvements are important in view of the increased number of repeaters in the circuit.

The use of this improved type circuit has been extended during the last few years to connect a large number of the important cities in the United States. A few of the longer circuits are:

Circuit	Approximate Length of Circuit Statute Miles	No. of Through Line Repeaters
Boston-Chicago	1,180	4
Chicago-Denver	1,090	3
Chicago-Los Angeles	2,890	12
Chicago-San Francisco	2,410	8
Dallas-St. Louis	670	2
Denver-San Francisco	1,350	4
Jacksonville-Havana	640	3
New York-Chicago	940	3
New York-Havana	1,710	8
New York-New Orleans	1,100	6
New York-St. Louis	1,020	5
Kansas City-Denver	750	2

One of the most recently established of these circuits is the Chicago-Los Angeles circuit routed over the southern transcontinental line. This and other through circuits on this line from Denver via El Paso to Los Angeles were established last year in order to provide for the growth in transcontinental traffic and to make available a second route as protection for the through service to the Pacific Coast. A brief description of these circuits will be given as typical of the long open wire circuits on this and other routes.

In the following, certain data based on actual experience with circuits on the southern transcontinental route, and in certain instances on circuits on the central transcontinental route are given. These data, however, are in general representative of results obtained on circuits of the same type throughout the Bell System.

From Denver west, a phantom group of four 165 mil wires provides for a Chicago-Los Angeles circuit, a Denver-El Paso circuit,

[3] Refer to paper entitled "Practical Application of Carrier Telephone and Telegraph in the Bell System," by Arthur E. Rose, *Technical Journal*, April, 1923.

an El Paso-Los Angeles circuit and another circuit between Denver and Los Angeles with stations at intermediate points along the line. East of Denver facilities of the same type on an existing through route via Kansas City to Chicago with four intermediate repeater stations are used for the Chicago-Denver portion of the Chicago-Los angeles circuit. The facilities and equipment arrangements permit rapid changes at the various repeater stations so that the circuit layout may easily be changed to take care of temporary rearrangements necessitated by trouble and to set up different layouts for the evening and night loads which at present are heavier than the day load.

Duplex telegraph equipment has been installed at various stations along the new route from Denver to Los Angeles for operating four direct current telegraph circuits derived by compositing the open wire circuits. In addition, a 10-channel carrier current telegraph system has been installed. Thus a total of 17 circuits, 3 telephone and 14 telegraph, operating on four wires are at present available over the new route for the through service. This requires a considerable

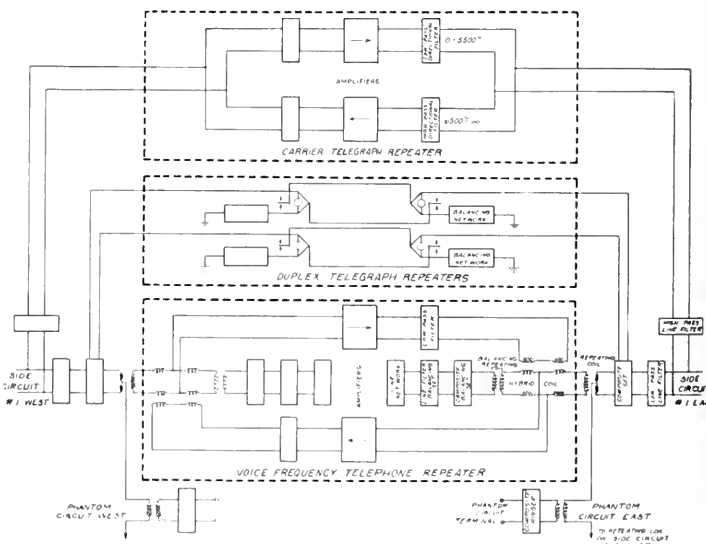


Fig. 2 Simplified Equipment Layout Intermediate Repeater Station Chicago-Los Angeles Circuit

amount of equipment at each of the repeater stations as illustrated by Fig. 2 which shows a simplified equipment layout at a typical repeater station. In addition to this number of circuits, a carrier current telephone system or an additional carrier current telegraph system can be provided and operated over these same wires at such time as traffic growth may warrant.

In view of the importance and number of services routed over the through wires, careful consideration was given to construction features. Copper line wire 165 mils in diameter affording high mechanical strength as well as low transmission loss was provided. For a

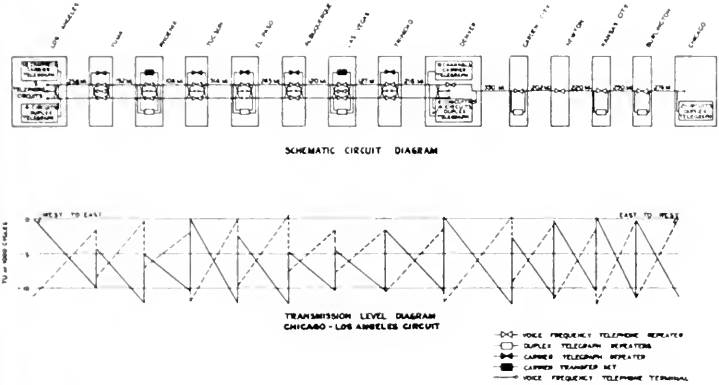


Fig. 3

large part of the distance an existing pole line over this route was used, but a new line was constructed in certain sections where the existing line was carrying full capacity load or was otherwise not suitable, and in other sections to avoid, as far as practicable, the use of intermediate or toll entrance cable. In a number of cases an important consideration in rerouting was the avoidance of exposures to electric light and power circuits. Considerable pole replacement work also was done in some of the existing pole line sections to strengthen the structure. The wires were transposed in accordance with a design[1] providing low crosstalk values at the frequencies used in carrier operation as well as within the voice range.

In determining the location of repeater stations for the telephone circuits, direct current telegraph and carrier systems, it was neces-

[1] "Carrier Current Telephony and Telegraphy," J. H. Colpitts and O. B. Blackwell, *Journal A.I.E.E.*, Vol. XL, No. 5, May, 1921.

sary to consider the economical limit of attenuation loss in each repeater section, the degree of balance to be obtained between the line impedances of the different sections and the corresponding balancing network impedances, and the proper transmission levels, as well as the limited choice of points where it would be practicable to maintain these stations from an economy and maintenance force standpoint. Fig. 3 shows the layout of through circuits on the southern transcontinental line and a transmission level diagram of the Chicago-Los Angeles telephone circuit, indicating the location and spacing of repeater points, attenuation losses in the different sections and amplification of repeaters.

Impedance Characteristics. It has been practicable in the construction of the new facilities to avoid long sections of intermediate or entrance cable except at a few points and in general, very smooth impedance characteristics of the different repeater sections have been obtained. At the points where appreciable lengths of cable could not well be avoided, a special type of loading[5] has been designed

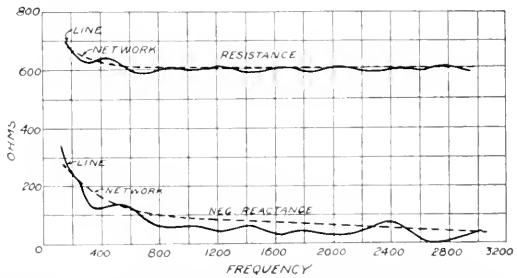


Fig. 4—Impedance Characteristic 216 Mile Repeater Section Non-Loaded 165 mil Physical Circuit (Circuit Terminated to Appear as Infinite Line)

for the purpose of raising the impedance of the cable circuits to values that match the impedance of the open wire at the carrier frequencies as well as at voice frequencies. This loading also is of particular benefit in reducing the attenuation loss at carrier frequencies which in non-loaded cable may be comparatively high.

The impedance characteristic of a typical repeater section 216 miles long is shown by the heavy lines in Fig. 1. The circuit in this case is terminated at the distant end by a network which makes it

[5] Refer to paper on "Carrier Current Telephony and Telegraphy," by Colpitts and Blackwell, previously noted.

appear as an infinitely long line. The slight irregularity indicated by the humps in the impedance curve is due to a short section of non-loaded entrance cable at the distant end. Fig. 5 shows the same section of line terminated at the distant end by the impedance of the repeater into which it normally works. The impedance character-

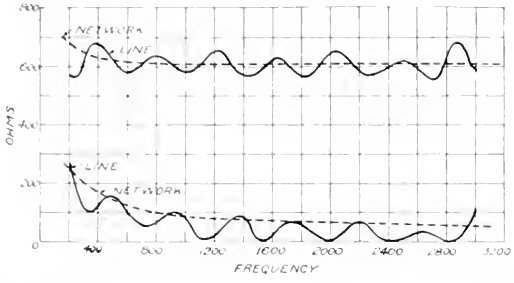


Fig. 5—Impedance Characteristic 216 Mile Repeater Section Non-Loaded 165 mil Physical Circuit Terminated at Distant End by Passive Impedance of Adjacent Repeater

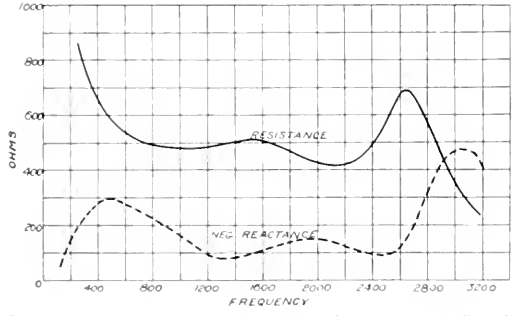


Fig. 6—Passive Input Impedance Characteristic of Improved "22" Type Repeater

istic of the repeater is shown in Fig. 6. The dotted curves in Figs. 4 and 5 are the impedance components of the network used to balance the line circuit. This network is of the precision type,[6] designed for use in connection with long non-loaded open wire circuits which employ repeaters amplifying a wide band of frequencies.

[6] "Telephone Repeaters," by B. Gherardi and F. B. Jewett, A.I.E.E. Transactions, Vol. XXXVIII, No. 11, November, 1919. See also "Impedance of Smooth Lines and Design of Simulating Networks," Ray S. Hoyt, *Technical Journal*, April, 1923.

Transmission Characteristics of Line and Repeater. Fig. 7 shows the attenuation frequency characteristic of a typical repeater section expressed in *TU*.[7] The amplification frequency characteristic of the telephone repeater shown in Fig. 8 is such as to compensate for the inverse characteristic of the line circuit so that the over-all transmission characteristic of the circuit will be uniform over the important frequencies of the speech range as illustrated by Fig. 9.

Signaling. On the shorter of these circuits employing only a few repeaters, signaling current is relayed at each repeater point, new energy being sent into the adjacent section of the line by the operation of relays associated with the repeater. 135-cycle current is used for the signaling current sent over the line, this being the frequency commonly used for signaling over composited circuits. On longer circuits employing several repeaters, the time lag of the ring can be decreased by a system employing a combination of amplified and relayed ringing at alternate repeater stations. At points where the ring is amplified, it is necessary to increase the repeater amplification at 135 cycles in order that sufficient ringing energy may reach the relaying repeater point to operate the ringing relays. This is accomplished by making slight changes in the input circuit of the repeater to increase its efficiency at the lower frequencies as illustrated by the dotted curve in Fig. 8. Best results are obtained by relaying at alternate repeater points.

At each relayed ringing point a certain time interval is required for the operation of the relays and for this reason the length of time during which the ringing current is applied to the line may become less and less for each succeeding repeater. If the ring, therefore, is not of sufficient duration, it is likely that sufficient ringing energy to operate the line signal will not be received at the distant terminal. This has introduced some operating difficulties and made it necessary to exert great care in the maintenance of the apparatus at the intermediate as well as at the terminal stations involved and careful overall checking and lining up of the circuit as a whole.

There has been developed a system employing signaling currents of voice frequency which has largely overcome these difficulties. The signaling current is amplified by the repeaters with approximately the same efficiency as the voice currents so that relaying is unnecessary. Particular attention has been given in the design of the system to preventing false operation of the signals from voice or extraneous currents.

[7] See article in this issue "The Transmission Unit and Telephone Transmission Reference System," by W. H. Martin.

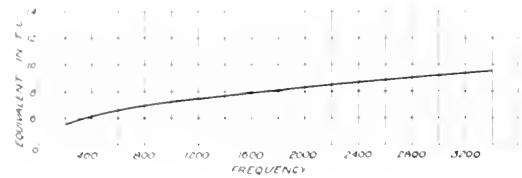


Fig. 7—Transmission-Frequency Characteristic 216 Mile Repeater Section of Non-Loaded 165 mil Physical Circuit

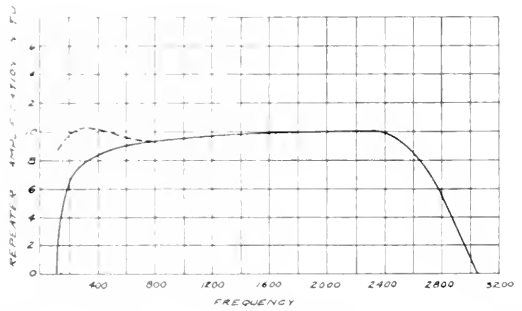


Fig. 8—Amplification-Frequency Characteristic of Improved "22" Type Repeater

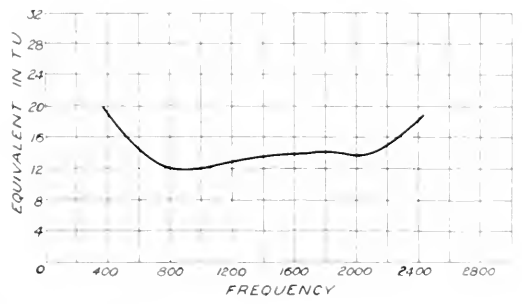


Fig. 9—Overall Transmission-Frequency Characteristic of Long Non-Loaded 165 mil Circuit Denver to Los Angeles

At the present time this system of voice frequency signaling is employed for regular use on the Chicago-Los Angeles circuit and a few others. The system employing 135-cycle current alternately relayed and amplified at repeater points also is installed on these circuits and probably will be retained for emergency use and to permit temporary changes in circuit layout.

Maintenance. The continuity of the many important services routed over the facilities used in making up these long circuits is dependent upon continuous and efficient maintenance methods and performance. Coordination of the work of the different offices is most essential in order to obtain best results, especially on the longer direct circuits and on those built up by the connecting together of several circuits, as there are a large number of variable factors. To assist in obtaining best results, accurate records of the circuit make-up

TOLL CIRCUIT LAYOUT RECORD

CIRCUIT NO. 1 Chicago (MF) Los Angeles EQUIVALENT 12.9 CIRCUIT ORDER 5617 ITEM 3
COMPUTED DATE IN SERVICE
CONTROL OFFICE Chicago CLASSIFICATION VL MEASURED 12 CARD ISSUE NO. 1 DATE 4-17-24

	FROM	TO	CABLE OR LINE	PAIRS OR PINS	SIZE OF WIRE	LOADING	LENGTH	EQUIV.	REPEATING COILS ON SIDE	ON PHANT.	ON PHANT.	FOR 4W TERM.	CX	RINGER	OTHER	TOTAL LOSS (Volumes) TP W
C	(Chicago	Morl Pk.	1 Chge-MP Ca	127	16	M	7.60	1.4							BL	80
D	Morl. Pk.	P7415	Chicago-Omaha	9-10	165	N	175.3	6.1		55D			TS		KK	1.05
E	P7415	P7445-1/2	" - "	9-10	165	N	.7	.0								
F	P7445-1/2	Davenpt.	Cable	7	13	N	.06	.0							KK	.10
G	Davenpt.	P7445-1/2	"	63	13	N	.06	.0								
H	P7445-1/2	P7557	Chicago-Omaha	35-36	165	N	2.1	.1								
J	P7557	P7741-1/2	" "	25-26	165	N	4.5	.2								
K	P12976	P9047-1/2	St. L.-Dav.	5-6	165	N	83.4	2.9								
L	P9047-1/2	Burlington	Cable	5	10	N	.51	.2		75A			TS			.90
M	Burlington	P9047	"	15	13	N	1.03	.5		75A			IS			.90
N	P9047	P7903-1/2	St. Louis-Dav.	15-16	165	N	39.5	1.0								
P	P1	P612	Burl.-Kan.City	5-6	165	N	14.7	.5								
Q	P612	P1578	" " "	5-6	165	N	83.0	.8								
R	P1578	P5634	" " "	5-6	165	N	96.2	3.4								
S							TOTAL		Required Equiv. 12.9						TOTAL	

TELEPHONE REPEATER DATA

STATION	CLASS	RINGING	OFF	TOWARD A: TRANS. FREQ.	TOWARD A: REC. FREQ.	TOWARD A: REC.	TOWARD A: B.O.B.	TOWARD A: TRANS. POWER	TOWARD B: TRANS. FREQ.	TOWARD B: REC. FREQ.	TOWARD B: REC.	TOWARD B: B.O.B.	TOWARD B: TRANS. POWER	RINGING ON 1 W SIDE OF 4 W TERM. EQ	DISTRIBUTION: OFFICE	BR	REP	TM	REPTR
T				Req.				Req.	Req.				Req.		Morl.Pk.		A		
U	Morl. Pk.	CC	M					3			17B		25		Burlngtn.		•		
V	Burlgtn.	TLI	Amp.	N	11.8		17F	BL	25	11.8		17B	BL	25	Omaha		•	•	
W	Kans.City	TLI	135	N	11.8		17F	BU	25	11.9		17B	BU	25	Kans.City		•	•	
X	Denver	TLI	Amp.	N	11.8		17F		21	11.8		17B		25	(Over)				

Fig. 10

from end to end, including a complete description of the types of equipment and transmission data are prepared and furnished to the terminal and repeater stations. These are made on cards of convenient size, as illustrated by Fig. 10, which is one of the five cards for the Chicago-Los Angeles circuit. To insure proper functioning of the circuit and satisfactory overall transmission and signaling one of the terminal offices of each circuit is designated as the controlling office for that circuit and is responsible for the direction and super-

vision of tests and adjustments required on the circuit as a whole. In addition to the duties in connection with the maintenance of the circuit as a whole each office along the circuit is responsible, of course, for the proper physical maintenance of the plant in its territory.

High grade maintenance is necessary to reduce to a minimum, service interruptions, noise and crosstalk and fluctuations in circuit characteristics and equivalents. An important part of this work consists of frequent periodic inspections, measurements[5] of insulation resistance, loop resistance, resistance balance, transmission, noise and crosstalk and equipment parts which are subject to variation.

In order to make many of the measurements and tests it is necessary to remove the circuit from service. This would result in considerable lost circuit time if each of the stations made such measurements and tests independently. In order to minimize this lost circuit time, therefore, it has been found desirable in the case of long telephone circuits of this type to institute what is known as "co-ordinated testing" procedure. Under this procedure a definite time is set aside for the periodic tests and all repeater stations and both terminal stations co-ordinate their work under the direction of the controlling office. The success of this system is dependent upon each station doing its part of the work correctly and within a specified time allowed for each test. The method of conducting the tests is illustrated in the following description.

1. *Roll Call*—The tester at the controlling office first calls the roll, starting with the first station and proceeding through to the distant terminal, each station replying by name and giving the temperature and weather conditions.

2. *Repeater Amplification and Vacuum Tube Tests*—The tester at each station measures the amplification in both directions given by the telephone repeater at that point and checks the condition of the vacuum tubes.

3. *Balance Tests*—At each repeater station the degree of balance between the line circuit and the balancing network circuit is checked in both directions. Since it is necessary that each section of the circuit be terminated at the opposite end from the station making the balance tests, alternate repeater stations terminate the circuits and the other stations proceed with their balance measurements. The procedure is then reversed.

[5] For description of these tests and their application see article in this issue "Electrical Tests and Their Applications in the Maintenance of Telephone Transmission," by W. H. Harden.

4. *Transmisson Equivalent*—When the balance tests have been completed, a measurement of the overall transmission loss is made between the terminal stations.
5. *Talking Test*—In order that the quality and volume of transmission from a service standpoint may be determined, a talking test is made over the entire circuit using standard subscriber sets at each end.
6. *Signaling*—As a final check, ringing tests are made over the circuit in both directions to insure that satisfactory signaling is being obtained.

This testing routine has been perfected to such an extent that the circuit need not be kept out of service for more than about 15 minutes even in the case of the longest circuits. Results of measurements over the period of a year on the Chicago-San Francisco circuit are shown in Fig. 11. The overall transmission measurements, which

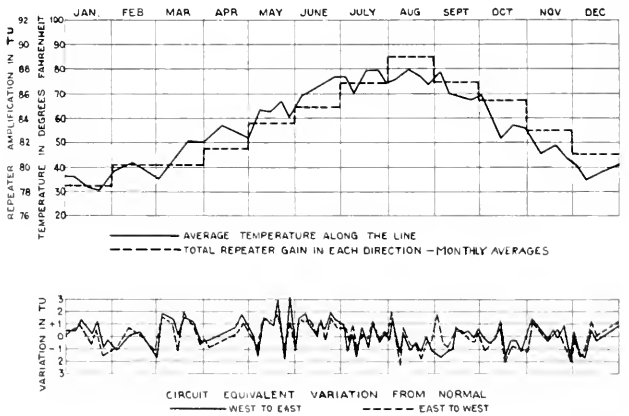


Fig. 11—Monthly Ranges in Temperature and Repeater Gain on the Chicago-San Francisco Circuit

are shown as variations from normal, were made at the conclusion of semi-weekly tests and after any necessary adjustment in repeater amplification had been made to compensate for changes in attenuation loss. The other curves show the average temperature along

the line at the time of tests and the average total repeater amplification from which can be noted the amount of amplification required to offset the variation in transmission equivalent due to seasonal temperature changes.

Conclusion.—As mentioned earlier in this paper the data and results given in the foregoing, although applying particularly to circuits on the southern and central transcontinental routes, are also representative of the conditions on other long circuits of the same type. The establishment of these high quality circuits, which are also available for the application of carrier systems and which in certain cases have been so equipped, constitutes another important step in bringing together all sections of the country by telephone.

Vacuum Tube Oscillators—A Graphical Method of Analysis

By J. W. HORTON

Introduction

THE vacuum tube oscillator is fast becoming one of our most versatile circuits and the requirements which are being imposed upon it are constantly increasing in severity. In some cases it is asked to efficiently convert several kilowatts of direct current power to alternating current power. At other times, it may be called upon to deliver an alternating current having a frequency which shall remain constant within extremely narrow limits. It may be required to operate at a few cycles per second or at several million.

The question of frequency stability has recently taken on considerable importance. The need for currents of accurately known frequency is being felt in all branches of the electrical communication art, particularly in the field of multiplex transmission over wires by means of carrier currents and in radio broadcasting. The factors affecting the frequency of an oscillator will for this reason be given attention in the following discussion.

The operation of a vacuum tube oscillator or, in fact, of any system maintained in continuous oscillation, has certain unique features. In order for such a system to be in stable equilibrium its several elements must adjust themselves until certain necessary conditions are established. It is important, in an analytical study of oscillators, to know the manner in which this adjustment takes place.

If any operating condition may be defined by an equation made up of independent variables, it is a relatively simple matter to predict the result of changes in a single one. When, however, a change in one quantity is accompanied by a general readjustment of all the others, it is quite difficult to obtain a clear picture of what occurs from an equation. Graphical methods are better suited to a study of the manner in which a number of inter-dependent variables arrive at an equilibrium condition. Such a graphical treatment will be described in the following paragraphs and its application to the design of a circuit to perform certain specified duties will be discussed.

Graphical Method for Determining Conditions of Stable Operation

It is sometimes convenient to think of an electrical transmission system as being made up of a number of units, each delivering energy

to the next succeeding unit, and thereby controlling the energy which that unit, in turn, delivers to the next. In case such a unit is made up of a vacuum tube amplifier circuit with its associated power supply batteries, it will be capable of passing on to succeeding units a greater amount of energy in a given time than it receives from preceding units. If a transmission unit does not contain some source of energy, it will, in general, deliver less power than it receives. In many cases these units may be arranged so as to form a complicated network. Whenever in such a network, a group of units forms a closed loop, that particular group is said to constitute a regenerative system. If a regenerative system is capable of maintaining a continuous flow of energy around the loop without receiving energy from any unit

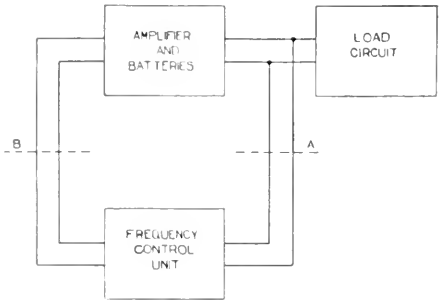


Fig. 1. Elements of an oscillating system

of the transmission network external to the loop, the system is said to be oscillatory.

For the purpose of this discussion let us think of an oscillatory system as made up of three units, the amplifier with its associated power supply source, a frequency control unit and an energy absorbing load unit. The arrangement of these units is as shown in Fig. 1. Now it is quite possible to determine the individual characteristics of the amplifier and of the frequency control units considered separately. The problem is to find the relation between these individual characteristics and the characteristics of the system.

In order that the regenerative circuit shall be in stable equilibrium, there are two conditions which must be met. The first of these is that the increase in power from the point B to the point A, through the amplifier unit, must be exactly equal to the decrease in power from the

point A to the point B, through the frequency control unit. Due account must be taken of any energy delivered to the load circuit. In other words, when a given amount of energy flows into the amplifier across the junction B, it must be transmitted around the regenerative loop and returned to this junction unchanged in amount. The second condition is that the phase displacement of the wave transmitted from A to B through the frequency control unit must be equal in amount and opposite in sign to the phase displacement of the wave transmitted from B to A through the amplifier. That is, a wave which enters the amplifier at the junction B must be transmitted through the regenerative system and returned to this junction with no resultant phase displacement.[1] The individual characteristics of the amplifier and of the frequency control unit which permit these two conditions to be satisfied fix the operating point of the system.

Although the reasoning to be used in the succeeding paragraphs may, in general, be applied to any oscillatory system, it will be easier to follow if described in terms of familiar electrical circuits. In Fig. 2

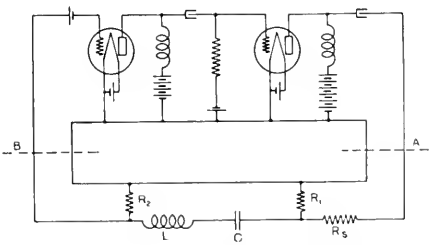


Fig. 2 — Circuit of an elementary oscillator

the amplifier and associated batteries are shown in the form of an elementary vacuum tube circuit. It is necessary to use a two-stage circuit if the voltages at A and at B are to have the same sign. This is because the grid voltage, which is obtained as a potential drop due to current from an external source flowing through a resistance connected between the grid and the filament of the vacuum tube, reduces the current flowing from the filament of the tube to the plate, through a second external resistance, as the current flowing from the grid to the filament is reduced. That is, a change in the voltage drop across

[1] This condition for stable equilibrium will also be satisfied if the total phase shift around the loop is equal to $2\pi n$ where n may be any whole number.

the grid circuit resistance causes a change of opposite sign in the voltage drop across the plate circuit resistance, the two voltage drops being referred to the potential of the filament. The frequency control unit is in the form of a series circuit containing inductance, capacity, and resistance. Two resistance elements are used for coupling to the input and to the output of the amplifier.

Let us first consider the properties of the vacuum tube amplifier. In Fig. 3 the voltage developed across the junction A is plotted as a function of the voltage across the junction B. Let us assume, for the present at least, that this curve holds for all frequencies. Obviously

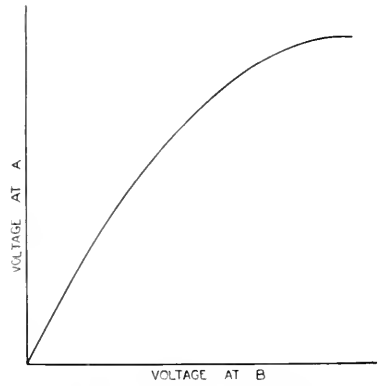


Fig. 3.—Amplifier output characteristic

the voltage across the junction A depends upon the impedance looking into the frequency control unit, but if the resistance R_1 is small in comparison with the resistance R the voltage will be practically independent of the frequency control unit. This curve represents a familiar characteristic of the vacuum tube amplifier. It shows that as the voltage upon the grid of the first tube is increased, a point is reached where the amplitude of the output is no longer proportional to the amplitude of the input. If this is carried far enough a point is ultimately reached where a continued increase in the voltage on the grid fails to produce any further increase in the voltage across the output. For our present purpose the data contained in this curve will be more useful if plotted in a less familiar form.

In Fig. 4 the ratio of the voltage across the junction B to the voltage across the junction A is plotted as a function of the voltage across the junction A. This curve is obtained from the same data as the curve of Fig. 3 and tells the same story. Assuming that this curve holds for all frequencies a family of curves may be plotted for the amplifier unit, as shown by the horizontal lines in Fig. 5. In these

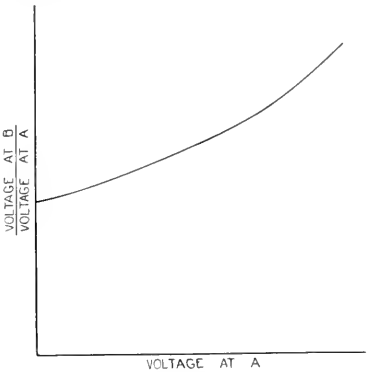


Fig. 4—Amplifier gain characteristic

curves the ratio of the voltage received by the unit to the voltage delivered by it is plotted against frequency. The numbers associated with each curve indicate the voltage at A, in arbitrary units, for which the curve holds.

A similar family of curves may be plotted for the frequency control unit. Since the impedance of the series resonance circuit varies with frequency from relatively high values above and below the resonance frequency to a minimum value at the resonance frequency, it follows that, for a fixed voltage across the junction A, the current through the inductance, the capacity and the resistance R_2 will vary with frequency. Consequently the voltage drop across the resistance, which is impressed across the junction B, will vary with frequency. The relation between this voltage and frequency, for a fixed voltage across the junction A, is given by the familiar resonance curve. As the voltage across the junction A is increased, currents of considerable magnitude may be caused to flow through the inductance, particularly in the neighborhood of the resonance point. If this

inductance has an iron core an increase in the current will result in increased damping which, at a fixed frequency, acts to reduce the ratio between the voltage set up across the resistance R_2 and the voltage impressed on the junction A. The resonance curves given

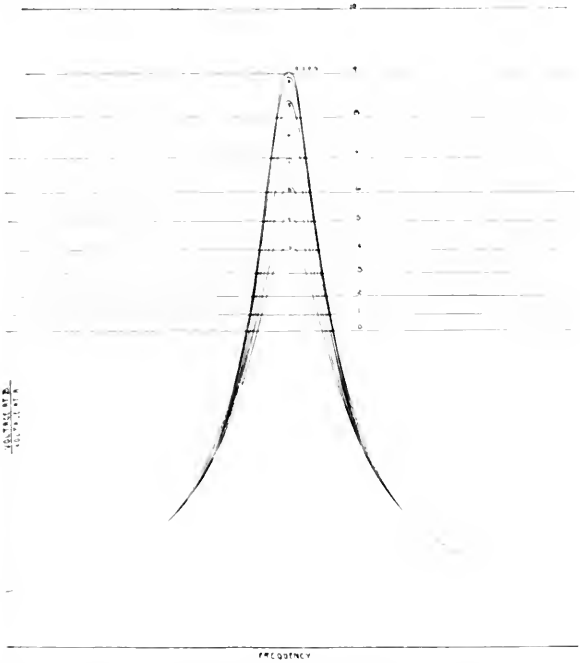


Fig. 5—Families showing relation between power gain, or loss and frequency for various power levels

in Fig. 5 show the relation between the ratio of the voltage across the junction B to the voltage impressed upon the junction A and frequency. The numbers again indicate the voltage at A, in the same arbitrary units as were used for the amplifier family.

Selecting any of the values given for the voltage across the junction A, it will be found that there are two curves showing the relation between the ratio of the voltage at the junction B to that at the

junction A and the frequency. One of these is a characteristic of the amplifier, the other of the frequency control unit. It is, of course, apparent that the voltage across any junction in a transmission system may be taken as a measure of the rate at which energy crosses this junction. Therefore, points of intersection of these lines satisfy the first condition which was imposed upon the oscillating system in order that it should be in stable equilibrium, namely, that the increase of power through one portion should be equal to the decrease in power through the remaining portion. Such points of intersection define values for the amplitude of the voltage at the junction A and of the frequency for which this condition is met. Similar pairs of lines, plotted for other values of the amplitude of the voltage at the junction A, have intersections indicating the corresponding frequency for which the energy relations are again satisfied. For each of these points, then, the amplitude of the voltage at A, the frequency and the ratio of the amplitude at B to the amplitude at A have the same values for the amplifier unit that they have for the frequency control unit. In the curve A, of Fig. 6, the first of these

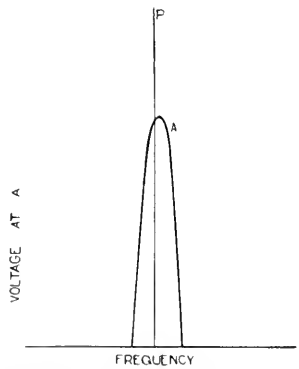


Fig. 6 — Amplitude and phase equilibrium curves

variables is plotted against the second. The curve, therefore, shows the magnitude of the voltage delivered by the amplifier unit which, for a given frequency of the wave transmitted by the system, permits energy equilibrium to be maintained.

If energy considerations alone determined the stability of the oscillating system, it would appear that the operating point might be anywhere along this line. It is necessary, however, to consider the phase displacements occurring in the two units as well. The phase difference between the voltage across the output resistance of the frequency control unit and the voltage impressed across the junction A varies with frequency as indicated by the family of curves shown in Fig. 7. The assumption is made in drawing these curves

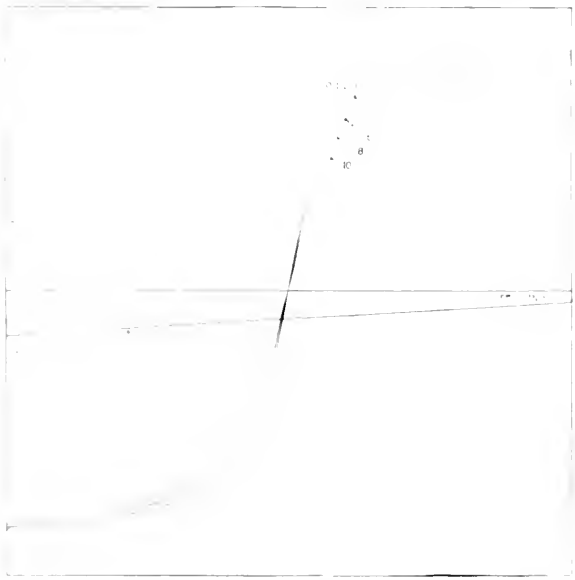


Fig. 7. Families showing relation between phase displacement and frequency for various power levels

that the change in damping is due entirely to a change in the resistance of the coil and that the inductance and the natural frequency of the circuit are unaltered as the load is increased. For low damping in the resonant circuit the phase shift changes rapidly with frequency in the neighborhood of the resonance point. As the amplitude of the voltage across the junction A is increased, thereby

increasing the damping, the rate of change of this phase shift is reduced. The several curves correspond to different values of the voltage amplitude at the point A.

The phase relation between the voltage wave impressed upon the input to the amplifier and the wave delivered by it is indicated in Fig. 7 by the single straight line. That is, we are assuming that the phase displacement of the wave transmitted through the amplifier varies but little over the frequency range covered by the diagram and that it is independent of the power which is being delivered by the amplifier. The numbers associated with the several curves have the same significance as those used with the power ratio families.

From these two sets of lines it is possible to determine a series of values of the voltage at the junction A and of the frequency for which the resultant phase displacement around the loop is zero; exactly as we determined a series of values for which the resultant power change was zero. These values are plotted in Fig. 6 as shown by the line P. If the condition for zero phase shift were the only one which the system had to satisfy, it is obvious that it would be in equilibrium at any point on this curve. Since, however, the system is called upon to satisfy two conditions, one defined by the curve A and the other by the curve P, any intersections which they may have are the only points at which the system can be in equilibrium.

This method of analyzing the relations between the characteristics of the several members of an oscillating system, and their mutual adjustment to an equilibrium condition may be summarized in general terms. The system is considered as a regenerative transmission circuit divided into two portions. For each of these portions a family of curves is plotted showing the relation between the rate at which energy crosses one of the junctions, which will be used as a reference point, the ratio between the rates at which energy crosses the two junctions and the frequency. Any two of these variables may be chosen as the coördinates for these families of curves, the remaining variable being the parameter. The intersections of a curve in one family with the curve of the same parameter in the other family define pairs of values of the frequency and of the power at the reference junction for which the system is in energy equilibrium.

For each portion of the regenerative system, a second family of curves is plotted showing the relation between the power at the reference junction, the phase displacement of the transmitted wave between the two junctions and frequency. Intersections of a curve in one of these families with the curve having the same parameter

in the other family define pairs of values of the frequency and of the power at the reference junction for which the system is in phase equilibrium.

The relations between these two quantities—frequency and power—may be expressed by two curves, one indicating the values necessary for energy equilibrium, the other indicating the values necessary for phase equilibrium. The intersections of these curves correspond to the only values meeting both conditions. The several elements must, therefore, adjust themselves to operate at the frequency and at the power indicated by such an intersection.

Effect of Variations in Circuit Elements

In addition to determining the frequency and power at which a given system is in stable equilibrium, it is important to be able to predict the effect upon these quantities of such changes as may be expected to occur in the elements composing the system. It is then a relatively simple matter to so redesign these elements that some particular effect shall be reduced, or increased, as desired. The circuit which has already been described may be used for illustrating the application of the graphical method in answering some of the questions occurring most frequently in connection with vacuum tube oscillators.

One of the more important problems concerns the reaction on the oscillating circuit of the load absorbing system. Let us imagine that an impedance, to which energy is to be supplied, is connected across the junction A. If this impedance is a complex quantity it will alter both the amplitude and the phase of the voltage across the junction. This will affect both families of curves (Figs. 5 and 7) which define the operation of the amplifier. If, for simplicity in the present discussion, we assume the load impedance to be a pure resistance, the major change will be a reduction in the voltage across A for a given voltage across B. The ratio of the voltage at B to that at A will be increased and the family of curves defining the power ratio relations between frequency and power ratio in the amplifier will thus be moved upward. The reaction upon the energy equilibrium curve (curve A, Fig. 6) will be to decrease both its height and its breadth. Assuming that the phase equilibrium curve remains unchanged, it is apparent that the frequency at which the system is in stable equilibrium will be increased and that the power delivered to the junction A will be decreased. Any change affecting the amplification of the vacuum tube circuit would react in much the same way.

Another question concerns the effect upon the amplitude at which the system operates as its frequency is altered by readjustment of the frequency control elements. If, for example, the capacity in the series resonant circuit is increased, the resonance curves of Fig. 5 will move to the left. Their shape also will be altered very slightly. Since the power ratio curves defining the operation of the amplifier are horizontal straight lines there will be a correspondingly slight change in the shape of the curve indicating the possible conditions for energy equilibrium. It will, of course, be displaced to the left by the same amount as are the resonance curves. If, however, the resistance of the resonant circuit varies directly with frequency, as it might through changes in hysteresis and eddy-current losses, the current through the series resonant circuit and through the resistance, R_2 will be increased. This increases the ratio of the voltage across B to the voltage across A and consequently lengthens the ordinates of the resonance curves shown in Fig. 5. The shapes of the curves will also be changed due to the change in the ratio of the reactance to resistance. Under these conditions the energy equilibrium curve, in addition to being moved to the left, will be increased both in height and in breadth.

The phase curves of the frequency control unit Fig. 7 will be moved to the left by the same amount as the resonance curves. Due to the slope of the phase family of the amplifier which we have assumed to be coincident straight lines, the intersections of the two phase families must move away from the point of zero phase displacement. The separation between the members of the phase family of the frequency control unit is greater here and consequently the phase equilibrium curve is less nearly vertical than before. The slope of the phase family of the amplifier also causes the phase equilibrium curve to move to the left by a slightly greater amount than the displacement of the resonance point of the tuned circuit. It is apparent, therefore, since the phase equilibrium curve moves farther than the amplitude equilibrium curve, that their intersection will move to a position corresponding to a lower value of the voltage at the junction A. This is true, of course, only if the change in the shape of the amplitude equilibrium curve due to the change in resistance of the inductance coil is small. It is also evident that the change in the frequency of the current delivered by the oscillator is greater than the change in the resonant frequency of the inductance and capacity.

These two examples are undoubtedly sufficient to demonstrate how a change in the constants of a single element of an oscillating system

necessitates a general readjustment of the other elements and how this readjustment reacts upon the operating point.

During the last few years the need for oscillating circuits of exceptionally high frequency stability has become more and more pressing. The requirements of multiplex telephony and telegraphy by means of carrier currents set particularly severe limits on the constancy of frequency of the alternating currents used. The efficient use of the ether in radio communication also places a very narrow tolerance upon any frequency variation in the carrier generators. It may be of interest, therefore, to consider some of the fundamental factors affecting the frequency stability of the vacuum tube oscillator.

Two lines of attack are open: we can design the several elements so as to reduce the possibility of a change in the value of their constants, or we can adjust the system so that unavoidable changes produce the least effect. It is in this latter connection that the graphical method of analysis is particularly helpful.

A change in the constants of any element of the oscillating system is going to result in a displacement or in a change in shape in at least one of the two equilibrium curves shown in Fig. 6. For a given change in either curve the horizontal displacement of their intersection will depend upon the slope of the other curve. The steeper one curve is, the less will be the frequency change resulting from any variation in the other. It, therefore, follows that we should make both curves as nearly vertical as possible.

Referring to the gain and loss families, Fig. 5, it will be seen that the slope of the amplitude equilibrium curve, and consequently the magnitude of the frequency change corresponding to a given change in the voltage at the reference junction, is determined by three things:

1. The separation between the lines defining the power gain in the amplifier; the less this separation, the less will be the frequency change accompanying a given change in the voltage.

2. The separation between the resonance curves defining the power loss in the frequency control unit; the less this separation, the less will be the frequency change accompanying a given change in the voltage.

3. The slope of the resonance curves; the steeper these curves, the less will be the frequency change accompanying a given change in voltage.

It appears then, that the change in frequency resulting from a given change in phase displacement, that is, accompanying any change in the phase equilibrium curve, may be reduced by operating

the vacuum tubes considerably below their overloading point, where the gain changes but little as the output is increased; by operating the tuned circuit at low power levels, where the damping, and consequently the loss, varies but little with changes in the input; and by keeping the damping as low as possible.

The slope of the amplifier gain family is, of course, a factor, but in practice it is found undesirable to permit the gain of the amplifier to vary with frequency. The slope of the phase equilibrium curve, which determines the change in frequency corresponding to a given change in transmission gain or loss, depends upon three things, as may be seen from Fig. 7. These are:

1. The distance from the point of zero phase displacement at which the phase family of the amplifier intersects the phase family of the frequency control unit; the less this distance, the less will be the frequency change accompanying a given voltage change.

2. The slope of the phase family of the frequency control unit; the more nearly vertical these curves are made, the less will be the frequency change accompanying a given voltage change.

3. The separation between the members of the phase family of the frequency control unit; the less this separation, the less will be the frequency change accompanying a given amplitude change.

If the phase family of the amplifier is not a single line, the separation between its members would be a factor. The slope of the curve also has a slight effect. The distance from the point of zero phase displacement, at which the two families intersect, may be reduced by reducing such reactive impedances as appear in the amplifier circuit. The slope of the frequency control unit family may be increased by reducing the damping of the tuned circuit. It may also be increased by reducing the phase displacement in the amplifier, thereby operating nearer the point of zero phase displacement where the rate of change of phase shift with frequency is greatest. The separation between the members of the phase family of the frequency control unit may be reduced by reducing the magnitude of such changes as occur in the damping. Moreover, since for various amounts of damping the several members of the phase family approach coincidence at the resonance point, it is again desirable to reduce any phase displacement of the amplifier in order to work as near this point as possible.

It has just been suggested that any reduction in the phase shift through the amplifier will make the phase equilibrium curve more nearly vertical. It will be noticed, however, that this causes the intersection between the phase equilibrium curve and the amplitude

equilibrium curve to occur in a portion of the latter where it approaches the horizontal. It would appear desirable, therefore, if frequency stability is to be pushed to the limit, to permit a slight phase displacement to occur in the amplifier in order that the intersection might be located at a place where the amplitude equilibrium curve is steeper. Any decrease in the slope of the phase equilibrium curve will be more than compensated for by the increase in slope of the amplitude curve. It will be apparent from the curves that such an adjustment reduces the amount of frequency change accompanying any change in phase displacement at the expense of amplitude stability. In practice, phase changes can be made smaller than transmission gain changes and we are consequently justified in placing most of the burden of holding the frequency to narrow limits on the phase equilibrium characteristic.

As a result of the foregoing analysis it appears that the amplifier should be designed so that, at the normal operating point, its gain varies but little with load and so that it introduces as small a phase shift as possible. The tuned circuit should have little damping and the variation in damping with load should be reduced to a minimum. Although these conclusions have been based upon the characteristics of a specific circuit, they apply equally well to other circuits of the same general form.

Design of Circuit for High Frequency Stability

The arrangement of an oscillating circuit embodying the features which the preceding section has shown to be essential, if the generated current is to be maintained within narrow frequency limits, is given in Fig. 8. The frequency control unit is a shunt resonant

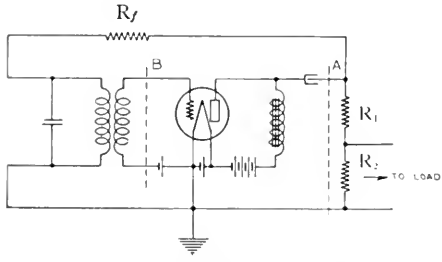


Fig. 8—Circuit of constant frequency oscillator

circuit coupled to the output of the amplifier, at the junction A, by a high series resistance, R_f, and to the input of the amplifier, at the junction B, through a winding coupled directly to the inductance.

The amplifier is designed to have ample load carrying capacity so that its gain varies but little with changes in load. This, as we have seen, is necessary in order to make the amplitude equilibrium curve steep and the frequency less subject to variation through unavoidable changes in phase displacement. Moreover, the voltage which appears at the junction A, as a result of a given voltage impressed upon the junction B, is stabilized by making the sum of the resistances R_1 and R_2 low as compared with the load impedance and with the impedance of the frequency control unit. In particular, R_2, across which the load is connected, is so small in comparison with the impedance of the load that changes in the latter are entirely negligible. Such an arrangement does not, of course, lead to high efficiency, but we must be prepared to make concessions in one direction in order to secure benefits in another. By making the effective load applied to the tube largely resistance the phase displacement occurring in the amplifier is made very small.

In this circuit it is not necessary to use two tubes to obtain the proper phase relations. At resonance the apparent impedance of the shunt resonant circuit approaches a pure resistance and the voltage drop across it is consequently in phase with the e.m.f. acting in the plate circuit of the tube. The current through the inductance, however, lags 90° behind this voltage. The e.m.f. set up in the oscillator input winding is 90° out of phase with the current in the primary, thus making it possible to secure in the frequency control unit the necessary 180° phase reversal between the plate and grid circuits of the tube. Care must, of course, be exercised to connect the windings of the oscillating coil in the proper direction.

The damping of the resonant circuit may be made small by giving the primary winding of the oscillating coil a high time constant. The coupling impedances introduce additional damping which may be made small by making both the feed-back resistance, R_f, and the input impedance of the tube high. The tube impedance may be made very high by using sufficient negative grid bias to prevent the filament-grid circuit from becoming conductive. Of the two coupling impedances, the feed-back resistance, together with the other impedances associated with the tube output, introduces the greater damping. Now it can be shown that for the frequency control unit to have a given transmission efficiency, the total added damping due to coupling is a minimum when the damping due to the input coupling

is equal to that due to the output coupling. It is, therefore, desirable to make the coupling to the input of the tube as efficient as possible in order to permit the coupling to the output to be reduced. For this reason the mutual impedance of the oscillating coil has been kept as high as is practicable.

By increasing the feed-back resistance the ratio of the voltage at B to the voltage at A may be reduced, thereby decreasing the ordinates of the power loss family of the frequency control unit Fig. 5. This affords a control by means of which the system may be adjusted so that both the amplifier and the frequency control unit are operated in regions where their power outputs are nearly porportional to the power inputs or, in other words, where the separation between the members in the gain and loss families is practically negligible.

There is another advantage in keeping the feed-back resistance high. In making it the major element in the network shunted across the resonant circuit, the effect of any variations in the output impedance of the tube or in the load impedance is reduced.

It is evident from an examination of the power ratio families which define the operation of the two elements of the regenerative system that before the system can come into equilibrium, at least one of these elements must enter a region where the relation between the power which it receives and the power which it delivers is non-linear. This means that the wave delivered by this element does not have the same form as the wave received by it. Distortion of this kind is manifest in the presence of harmonics of the fundamental frequencies in the current delivered by the oscillator. In most cases the amplifier is the distorting element and we find in the output all multiples of the fundamental. It has, however, been found advantageous in some instances to so adjust the system that the iron core of the inductance element in the frequency control unit overloads before the amplifier. In this case, the resulting distortion is such that only the odd multiples of the fundamental frequency are present.

By the proper choice of circuit elements, it has been found possible to design commercial oscillating circuits, covering the range of frequencies between 100 and 100,000 cycles per second, in which the frequency is but little affected by changes in elements external to the frequency control unit. In one such commercial oscillator it has been found, for example, that the average deviation in the frequency observed with any one tube from the mean frequency obtained with a number of tubes is approximately 0.02%. In this same circuit, as the plate potential changes from 100 to 150 volts, the frequency change does not exceed 0.04% at any portion of the fre-

quency range. Similarly, if the filament current is changed from 1.1 to 1.4 amperes, the average frequency change is 0.03%. Changes in the frequency resulting from changes in the load impedance are practically negligible.

Such frequency changes as occur in the oscillator referred to above, are due, to a large extent, to variations in the inductance consequent upon the variations in power level which accompany the particular circuit changes referred to. The stability of the system may, therefore, be increased considerably beyond the limits indicated if the electrical constants of the elements used in the frequency control unit are independent of the power level. The use of an air core coil in place of an iron core coil improves the stability to a very marked extent provided, of course, that the same time constant is obtained in both cases.

In oscillators which have been designed primarily for frequency stability, it is found that the largest frequency variation is due to the variation in the electrical constants of the frequency control unit with temperature. When iron core coils are used, the temperature coefficient of frequency of the oscillating system is approximately 0.01% for 1° C. Using suitably designed air core coils, the temperature coefficient of the oscillator becomes approximately 0.003% for 1° C. The change in frequency in this case is due almost entirely to the change in capacity of the mica condensers used in the frequency control unit.

Although the method of analysis which we have just considered has been discussed largely in terms of the relation of the frequency of an oscillating electrical system to the constants of the several members of the system, it is by no means limited to such consideration. It is, in fact, applicable to practically all types of oscillating systems, including those containing mechanically resonant devices. It should, however, be remembered that while an analytical study of this type may assist materially in furnishing a qualitative picture of the conditions existing in some piece of apparatus, it is by no means a substitute for a rigorous quantitative treatment.

Abstracts of Technical Papers[1]

Carrier Telephony on Power Lines.[2] N. H. Slaughter and W. V. Wolfe. The fundamental requirements of a telephone circuit are outlined briefly and translated into the terms of the power line carrier telephone problem. Considerable data on transmission line characteristics at carrier frequencies are presented, which clearly show the magnitude of the transmission problem and the best frequency values to employ. The advantages of using the "metallic circuit" arrangement rather than the commonly employed "ground return" arrangement are emphasized.

One of the chief problems in carrier telephony on power lines is to provide an efficient means of connecting the carrier equipment to the power line, and the various possibilities and preferred methods are discussed at some length.

The nature of the circuits and equipment employed are then described, together with an indication of their range of usefulness in power line telephone communication.

The Nature of Language.[3] R. L. Jones. In introduction, the history of human language is outlined and the manner of speech production is briefly described with special reference to English. Following this is a summary treatment of available data on the subject of speech and hearing. Much of this is the result of investigations carried out during the past few years in the Research Laboratories of the American Telephone and Telegraph Company and the Western Electric Company, at New York.

Human speech employs frequencies from a little below 100 cycles per second to above 6,000 cycles, a range of about six octaves. The ear can perceive sound waves ranging in pressure amplitude from less than 0.001 of a dyne to over 1,000 dynes and in frequency of vibration from about 20 cycles per second to about 20,000, a range of about ten octaves.

The intensities and frequencies used most in conversation are those located in the central part of the area of audition. The energy of speech is carried largely by frequencies below 1,000, but the characteristics which make it intelligible, are carried largely by frequencies above 1,000. Under quiet conditions good understanding is possi-

[1] The purpose of these abstracts is to supplement the contents of the *Journal* by reviewing papers from Bell System sources which relate directly to electrical communication but which will not be reprinted in the *Journal*.

[2] Journal A. I. E. E., Vol. XLIII, p. 377, Apr., 1924

[3] Journal A. I. E. E., Vol. XLIII, p. 321, Apr., 1924

ble with undistorted speech having an intensity anywhere from one hundred times greater, to a million times less than that at exit from the mouth. On the whole the sounds, *th*, *f*, *s*, and *v* are hardest to hear correctly and they account for over half the mistakes made in interpretation. Failure to perceive them correctly is principally due to their very weak energy although it is also to be noted that they have important components of very high frequency.

The Physical Criterion for Determining the Pitch of a Musical Tone.[4] HARVEY FLETCHER. This paper describes experiments in which a high quality telephone system was used to reproduce musical sounds from the voice, the piano, the violin, the clarinet and the organ without any appreciable distortion. Into this telephone system electrical filters were introduced which made it possible to eliminate any desired frequency range. Results with this system show that only the quality and not the pitch of such musical sounds changes when a group of either the low or high frequency components is eliminated. Even when the fundamental and first seven overtones were eliminated from the vowel *ah* sung at an ordinary pitch for a baritone, the pitch remained the same. These results were checked by a study of synthesized musical tones produced by ten vacuum tube oscillators, with frequencies from 100 to 1,000 at intervals of 100. It was found that three consecutive component frequencies were sufficient to give a clear musical tone of definite pitch corresponding to 100, and that in general when the adjacent components had a constant difference which was a common factor to all components a single musical tone of pitch equal to this common difference was obtained, but not otherwise. Recent work on hearing has shown that the transmission mechanism between the air and the inner ear has a non-linear response which accounts for the so-called subjective tones. When the components of low frequency are eliminated from the externally impressed musical tone, they are again introduced as subjective tones before the sound reaches the nerve terminals. Calculation of the magnitude of these subjective tones from the non-linear constants of the ear shows that the results on pitch are what might be expected.

Sound spectra of ten typical musical sounds, obtained with an electrical automatic harmonic analyzer to be described by Wegel and Moore, are given for *ah* sung at pitch *d*, *a* sung at *a*, piano c_1, piano e'', violin g'', clarinet c, organ, pipe, c_1 for three pressures, and organ pipe c'.

Ferromagnetism and Its Dependence Upon Chemical, Thermal and Mechanical Conditions.[5] L. W. MCKEEHAN. This review considers

[4] Physical Review, Vol. XXIII, No. 3, March, 1924.

[5] Journal of Franklin Institute, V. 196, pp. 583-601; 757-786, 1924.

first the general properties of ferromagnetic bodies and the particular forms of magnetization curves and hysteresis loops exhibited by iron, cobalt, nickel, and their alloys with each other and with other elements. The Heusler alloys are also described. The effects of temperature upon magnetization are then discussed in detail for the case of iron and the behavior of alloys is compared with this as a standard, both reversible and irreversible changes being discussed in some cases. The transient effects of mechanical strains within the elastic limit and the permanent effects of over-strain of the kinds usually met with in practice are considered. The review concludes with speculations in regard to the electronic groups in the atomic structure which are responsible for the occurrence of ferromagnetic properties. One hundred and forty references to recent periodical literature are intended to give starting points for more detailed study of any of the subjects discussed.

Permeater for Alternating Current Measurements at Small Magnetizing Forces.[6] G. A. KELLSALL. This is a description of a permeameter for making alternating current measurements of permeability on toroidal specimens at small magnetizing forces and at telephonic frequencies. It is a special type of transformer with a single turn secondary. The primary consists of a suitable number of turns of insulated copper wire wound directly on a finely divided toroidal magnetic core made of one of the high permeability permalloys. The single turn secondary is an annular copper shell enclosing the primary with an additional space provided for the core to be tested. The copper shell is provided with convenient means for opening and closing. The sample whose permeability is to be determined is interlinked with the open secondary which is then closed. The inductance of the instrument connected as one arm of an inductance bridge is then measured at the primary terminals. From the value thus obtained, the constants of the transformer and the dimensions of the sample, the permeability is computed.

Furnace Permeater for Alternating Current Measurements at Small Magnetizing Forces.[7] G. A. KELLSALL. This is an adaptation of the permeameter previously described for the measurement of permeability at elevated temperatures. It consists essentially of a permeameter with an addition of an annular electric furnace immediately surrounding the sample under test and suitably heat insulated from the other parts of the instrument. Like the simpler permeameter, it measures the permeability of ring samples for small magnetizing forces at

[6] J. O. S. A. and R. S. I., 8, pp. 329–338, 1924.

[7] J. O. S. A. and R. S. I., 8, pp. 669–674, 1924.

telephonic frequencies without the necessity of winding magnetizing coils upon them. The maximum temperature at which measurements can be made with this apparatus is about 1,000° C. By filling the unheated furnace with liquid air, a minimum temperature of —190° C is attainable making the whole range of the instrument about 1,200° C.

The changes introduced in order to adapt to permeameter for measurements at different temperatures do not impair its accuracy, the determination of permeability at both high and low temperature having the same precision as at room temperature.

Contributors to this Issue

W. H. Harden, B.E.E., University of Michigan, 1912; Engineering Department, American Telephone and Telegraph Company, 1912–1919; Department of Operation and Engineering, 1919–. Mr. Harden has been engaged in the development of transmission maintenance testing methods and in the preparation of routines and practices required for applying these methods in the telephone plant.

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.

W. H. Martin, A.B., Johns Hopkins University, 1909; S.B., Massachusetts Institute of Technology, 1911. American Telephone and Telegraph Company, Engineering Department, 1911–19; Department of Development and Research, 1919–. Mr. Martin's work has related particularly to loading, quality, and transmission of telephone sets and local circuits.

C. W. Smith, B.S.E., University of Michigan, 1916; Chicago Telephone Company, 1916–20; American Telephone and Telegraph Company, Engineering Department and Department of Operation and Engineering, 1920–.

Ray S. Hoyt, B.S. in electrical engineering, University of Wisconsin, 1905; Massachusetts Institute of Technology, 1906; M.S., Princeton, 1910. American Telephone and Telegraph Company, Engineering Department, 1906–07. Western Electric Company, Engineering Department, 1907–11; American Telephone and Telegraph Company, Engineering Department, 1911–19; Department of Development and Research, 1919–. Mr. Hoyt has made contributions to the theory of transmission lines and associated apparatus, and more recently to the theory of crosstalk and other interference.

KARL K. DARROW, S.B., University of Chicago, 1911; University of Paris, 1911-12; University of Berlin, 1912; Ph.D., in physics and mathematics, University of Chicago, 1917; Engineering Department, Western Electric Company, 1917—. At the Western Electric, Mr. Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

H. H. NANCE, Washington University, 1906-10; Long Lines Plant Department, 1910-23; district plant chief, 1917-18; division superintendent of equipment construction, 1919-20; division engineer, 1920-23; Long Lines Engineering Department, acting engineer of transmission, 1923; engineer of transmission, 1924.

J. W. HORTON, B.S., Massachusetts Institute of Technology, 1914; instructor in physics, 1914-16; Engineering Department of the Western Electric Company, 1916—. Mr. Horton has been closely connected with the development of apparatus for carrier current communication.

The Bell System Technical Journal

October, 1924

"The Stethophone," An Electrical Stethoscope

By H. A. FREDERICK and H. F. DODGE

I. Acoustic Stethoscopes

AUSCULTATION is commonly practiced by means of the ordinary stethoscope, a device with which the physician is able to study sounds produced within the heart, lungs, or other portions of the body and to determine whether such abnormal conditions exist as are evidenced by abnormal sounds. Of particular importance are the characteristics of the normal heart sounds, heart murmurs, breathing sounds and râles.[1] It is well known that the intensity of certain of these sounds is not in itself of fundamental significance, that, for example, certain very faint murmurs may represent serious organic lesions; hence it is of pathological importance that these sounds be heard and understood.

Most acoustic and mechanical vibratory systems introduce distortion by discriminating in favor of certain frequency bands. Extreme distortion may alter a sound beyond recognition. If a moderate amount of distortion is unavoidable, it may be possible to control it judiciously so as to give most accurate reproduction in the frequency region of major importance.

From this standpoint it is of interest to consider the frequency characteristics of the two common types of stethoscopes shown in Figs. 1 and 2. The stethoscopes used in these tests were equipped with thick-walled soft rubber tubing such that the distance from the chest piece to the ear pieces was approximately 55 cm. The characteristic of the open bell stethoscope was obtained by picking up sound from the surface of a piece of fresh beef and measuring the relative intensity of sound on a condenser transmitter[2] with and

[1] "The presence of any one of several types of lesions in or near the valves of the heart "gives rise to eddies in the blood current and thereby to the abnormal sounds to which we give the name murmurs." "No one of the various blowing, whistling, rolling, rumbling or piping noises to which the term refers, sounds anything like a 'murmur' in the ordinary sense of the word." R. C. Cabot Physical Diagnosis, pp. 182-3, 1923

"The term 'râles' is applied to sounds produced by the passage of air through bronchi windpipes which contain mucus or pus, or which are narrowed by swelling of their walls." R. C. Cabot Physical Diagnosis, p. 163 Râles may appear either as bubbling sounds, occurring singly or in showers, or as musical squeaks and groans.

[2] E. C. Wente, "The Sensitivity and Precision of the Electrostatic Transmitter for Measuring Sound Intensities," *Phys. Rev.* 19, No. 5, p. 498, 1922.

without the test stethoscope inserted in the sound path. In this experiment, it was impracticable to set up pure vibrations in the human body. A piece of fresh beef was a convenient substitute and one which for the purposes of such physical analysis appeared satisfactory.

The frequency characteristic of the open bell stethoscope is shown

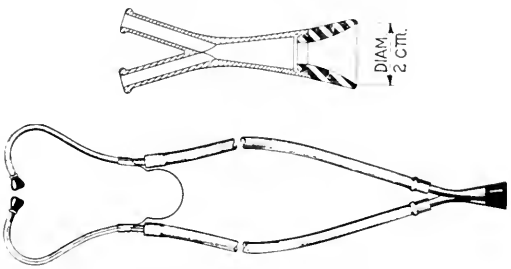


Fig. 1—The open bell stethoscope

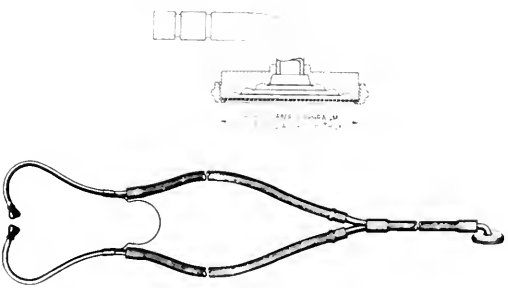


Fig. 2—One type of Bowles stethoscope

in Fig. 3, in which the "sensation value" as interpreted by the ear is plotted in transmission units,[3] *TU*, convenient units used to

[3] The transmission unit used in this paper is a logarithmic function of power ratio. The number of transmission units N corresponding to the ratio of two amounts of power P_1 and P_2 is given by the relation $N = 10 \log_{10} \frac{P_1}{P_2}$. The power ratio corresponding to N units is therefore $10^{\frac{N}{10}}$. For example, an increase of 10 *TU* signifies 10 times as much power; of 20 *TU*, 100 times as much power, etc. See W. H. Martin, "The Transmission Unit," *Journal A. I. E. E.*, Vol. 43, p. 504, 1924; *B. S. T. J.* Vol. 3, p. 400, 1924.

express relative loudness. A power ratio scale is also shown at the left and the power at 100 cycles is assumed equal to unity as a reference point.

This curve shows the relative efficiency of transmission for frequencies up to 2,000 cycles. The successive peaks are due primarily to resonance of the air columns and are partly determined by the length of the stethoscope tubing. Resonance thus increases the

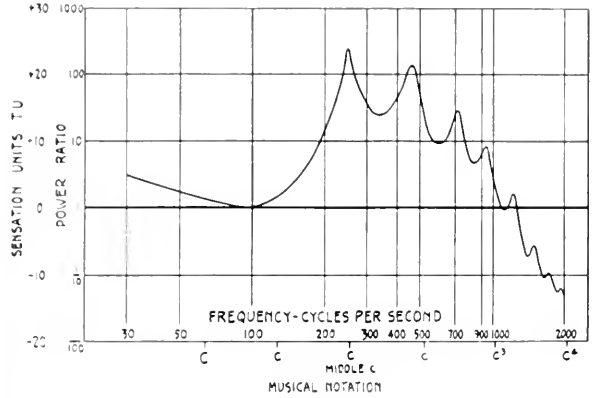


Fig. 3—Frequency characteristic of open bell stethoscope

efficiency of transmission at and above the fundamental peak frequency. As the frequency scale is ascended from this point, the transmission falls off gradually.

In a subsequent test, the open bell and Bowles types of stethoscopes (Figs. 1 and 2) were compared directly with one another. For this test, a vigorous sound was imparted to the sternum of a patient and the sound was picked up over the apex of the heart. Below 150 cycles, the Bowles stethoscope averaged approximately 15 *TU* less efficient, whereas, disregarding the somewhat different arrangement of the resonance peaks, between 300 and 1,000 cycles, it varied from 5 to 10 *TU* more efficient than the open bell type. These features of the Bowles stethoscope are due to the chest piece diaphragm. As will be shown in another paper, much of the energy of systolic and diastolic murmurs is made up of frequencies between 120 and 660 cycles per second. Thus, concurring with observations made

by Dr. R. C. Cabot,[4] it is to be expected that many of the moderately high and high pitched murmurs can be heard more distinctly with the Bowles than with the open bell stethoscope. On the other hand, for many faint pathological sounds such as presystolic murmurs which are composed primarily of relatively low frequencies, the open bell stethoscope is more satisfactory for observation. The latter introduces less distortion so that with it all sounds are retained more nearly at their original relative intensities. These remarks are, of course, confined to the particular designs of stethoscopes shown in Figs. 1 and 2. It should be noted that as the length of the rubber tubing is increased, the fundamental peak of Fig. 3 moves downward in frequency, and the transmission at higher frequencies becomes poorer. In order to retain the very high pitched components of certain heart and chest sounds, the use of long rubber tubing should, therefore, be avoided.

The common stethoscope serves as a convenient means of observing body sounds. If the available energy from a single chest piece is subdivided in order to supply several individuals, however, the sounds observed by each are much fainter. In cases where the sounds of pathological interest are sufficiently near the threshold of audibility the use of a multiplicity of observing tubes renders these sounds inaudible. This is often the case.

For teaching purposes or for consultation, it is extremely desirable to have multiple listening units. In the past, it has been necessary to handle the students of large classes either singly or in small groups. This method naturally limits the number of cases that can be demonstrated and makes it impossible to give each student as much practice as has been found necessary for him to become familiar with the more obscure sounds. Aside from these factors, it has not been feasible for a large group to observe simultaneously with the instructor the peculiarities and changes in murmurs of a transient or evanescent character.

With the development of vacuum tube amplifiers, the possibilities of reproducing and magnifying body sounds electrically were considered. It appeared that a device might be provided which would be useful not only in teaching but also in diagnosis, as an aid to physicians of subnormal hearing, in the reproduction of the very faint fetal heart sounds or even in fields beyond the scope of the ordinary stethoscope.

[4] R. C. Cabot, "Physicial Diagnosis," Chap. VI, 1923.

2. Early Development of the Electrical Stethoscope

The earliest development work on electrical stethoscopes was naturally centered about the carbon transmitter and other microphonic contact devices. In 1907, Einthoven [5] made records of normal heart sounds and murmurs. In 1910,[6] heart sounds were reproduced by a tuned mechanical relay consisting of a single microphonic contact and an electromagnetic element. With this device, heart sounds were transmitted audibly but evidently with a considerable amount of distortion, over a commercial telephone line in London. The normal heart sounds were amplified by Squier [7] for a group of physicians by means of a carbon transmitter in 1921. It is readily possible to amplify the fluctuations in current in a carbon microphone by means of vacuum tube amplifiers. However, the carbon microphone also introduces a certain amount of noise inherent in the use of loose contacts. This noise is below the threshold of audibility for the normal use of the microphone, as in the telephone plant, but when it is amplified along with the faint sounds of interest, in auscultation it becomes very annoying and tends to obscure these other sounds. This "microphone roar" contains components throughout the range of audible frequencies and hence cannot be eliminated. Various experimenters have, however, attempted to perfect such a device.[8,9] As far as we have been able to determine, such devices have not satisfactorily reproduced faint heart murmurs or chest sounds.

Of the other possible types, the electromagnetic has thus far appeared to offer the greatest promise. In design, this resembles closely the ordinary telephone receiver. This type requires a more powerful amplifier than the carbon microphone but this is not a serious limitation. Such a combination has been used with promising results to obtain graphical records of heart murmurs.[10] The progress made with this type of equipment for teaching purposes has been outlined.[11] The successful application of the electromagnetic

[5] W. Einthoven, "Die Registrierung der menschlichen Herztone mittels des Saitengalvanometers," *Arch. f.d. ges. Physiol.*, 117:461 April 1907; "Ein dritter Herzton," ibid. 120:31 Oct. 1907

[6] S. G. Brown, "A Telephone Relay," *Journal I. E. E.* May 5, 1910.

[7] S. W. Winters, "Diagnosis by Wireless," *Scient. Amer.* 124:465 June, 1921.

[8] R. B. Abbott, "Eliminating Interfering Sounds in a Telephone Transmitter Stethoscope," *Phys. Rev.* 21:200 Feb., 1923.

[9] Jacobsohn, "Amplified Audibility of Heart Sounds," *Berlin Letter J. A. M. A.*, 80:493 Feb. 17, 1923.

[10] H. B. Williams, "New Method for Graphic Study of Heart Murmurs," *Proc. Soc. Exper. Biol. and Med.*, 18:179 March 16, 1921.

[11] R. C. Cabot, "A Multiple Electrical Stethoscope for Teaching Purposes," *J. A. M. A.*, 81:298 July 28, 1923.

transmitter to teaching was due largely to the work of Dr. R. C. Cabot and Dr. C. J. Gamble at the Massachusetts General Hospital where a successful multiple electrical stethoscope was first employed for classroom lectures in June, 1923. The equipment consisted of an electromagnetic transmitter provided with a special form of mouthpiece for picking up the body sounds, a three-stage vacuum tube amplifier and a distribution system to accommodate as many as 125 students with single head receivers on which individual ordinary stethoscopes were held.[12]

The experience gained with this equipment indicated certain improvements to increase the sensitivity to body sounds, and at the same time decrease the disturbances caused by extraneous noises. Greater sensitivity required a better transference of sound energy from the body to the transmitter. Reduced room noise required that we couple the transmitter as closely as possible with the human body and at the same time make it insensititive to sound vibrations in the air. A preliminary analysis with electrical filters of the frequency characteristics of sounds of pathological interest to the physician showed that these sounds were composed largely of frequencies below 1,000 cycles. Inasmuch as the frequency characteristics of these various sounds are different, it has been found very useful to permit concentration on the sounds of interest by the use of electrical filters.

These factors led to the development of the electrical stethoscope called the "stethophone" which is described in the following paragraphs. This development was undertaken at the request and with the active cooperation of Dr. H. B. Williams of the College of Physicians and Surgeons, New York, Dr. Richard C. Cabot [11] of the Massachusetts General Hospital, Boston, and Dr. C. J. Gamble [12] of the School of Medicine of the University of Pennsylvania, Philadelphia. The cooperation of these physicians permitted the instrument to be given practical tests at every stage of its development.

3. General Description of the Stethophone

The stethophone consists essentially of the following elements:

1. An electromagnetic transmitter.
2. A three-stage amplifier with a potentiometer control.
3. A selected group of electric filters.
4. A multiplicity of output receivers for observers.

The whole is assembled in a substantial cabinet on wheels re-

[12] A detailed description of the apparatus used in this installation was presented in a recent paper. See Gamble and Replogle, "A Multiple Electric Stethoscope for Teaching," *J. A. M. A.*, Vol. 82, p. 387, 1924.

sembling a "tea-wagon." It requires for its operation a six-volt storage battery and a 130-volt "B" battery. These are housed in compartments in the lower part of the cabinet. Ten jack positions are provided to permit this number of persons to listen simultaneously

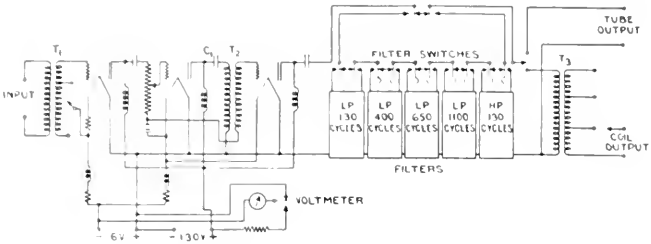


Fig. 4.—Circuit diagram of stethophone

around the stethophone. All controls are conveniently placed on a single panel to facilitate operation.

A schematic circuit diagram is shown in Fig. 4.

4. Transmitter

The transmitter employed with greatest success with the stethophone thus far is of the electro-magnetic type equipped with a special vibratory element which is placed in direct contact with the flesh of the patient.

One of the features of the transmitter is its insensitiveness to sound waves in the air. Thus, the ratio of extraneous noise picked up by the transmitter to the body sounds is greatly reduced so that observations can be made with a minimum amount of interference from room noise.

The transmitter construction provides efficient transfer of vibrational energy from the flesh or bony framework of the body to the vibratory steel element. It provides a means for coupling which serves as a mechanical transformer for body sound energy and avoids an abrupt change in the path of the waves and large attendant losses by reflection. The system is highly damped and minimizes the distortion of the sounds of interest.

Since the transmitter is a contact device, the physician may vary the pressure of application at will. Firm but light contact is desirable. The human flesh contributes damping to the vibratory system of the

transmitter. Undoubtedly this damping is not only variable for different individuals but depends upon the pressure and the nature of the flesh and bone structure in the vicinity of the point of application for any one individual. Thus the frequency characteristic of the transmitter is somewhat dependent on the conditions of use. The frequency of maximum response is slightly above 200 cycles, and the nature of the response-frequency curve indicates that the vibratory system is highly damped. A discussion of the overall frequency characteristic of the stethophone, including the transmitter, is given in a later section of the paper.

It is obvious that variations in the pressure of application will introduce disturbing noises in the audible frequency range. Suitable means have, therefore, been provided to eliminate the communication to the vibratory system of hand tremors, slight movements of the patient, and friction noises of the fingers on the case of the transmitter.

Another source of extraneous noise is the rubbing of the transmitter cord on the clothing or on other surfaces. A stiff cord is very objectionable from the standpoint of transmission of friction noises. Insulation from these noises has been provided by a very flexible section of cord at the transmitter end.

5. Amplifier

The three-stage amplifier employs one Western Electric 102-D and two Western Electric 101-D vacuum tubes. As shown in Fig. 4, the input transformer $T1$ connects the transmitter to the grid of the first tube which is coupled to the second tube through a resistance potentiometer. The second and third tubes are coupled through a transformer $T2$. The output circuit of the last tube may be connected to the load directly from its plate circuit for high impedance loads, or through an output transformer $T3$ for low impedance loads. The plate circuit of the second tube is tuned by means of a condenser $C1$ in order to retain high amplification at the low end of the frequency scale.

A very flat characteristic is obtained over the range of interest, the maximum variation being only about 3 TU (See Fig. 5). A total gain of about 80 TU is provided, that is, a power amplification of about one hundred million times. With an amplification of 50 TU, about the same loudness is observed in a single receiver in the output circuit of the stethophone as is heard by the direct use of the open bell stethoscope. This leaves a reserve amplification

of about 30 *TU* available for obtaining greater intensity of sounds or for supplying a large number of individual listening units.

The potentiometer between the first and second tubes makes it possible to adjust the amplification in small steps, each step giving

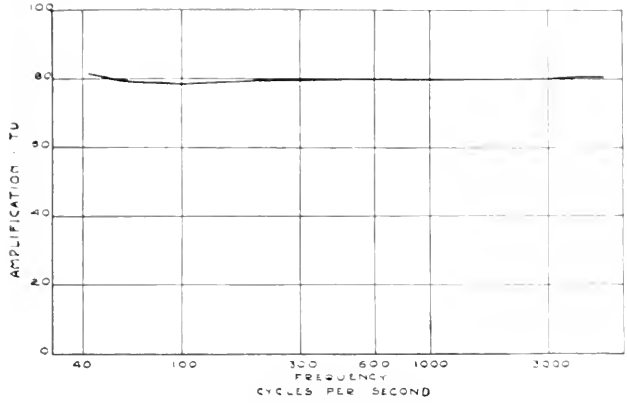


Fig. 5. Amplifier characteristics—maximum amplification

approximately twice the energy of the preceding one. This is an essential element of a flexible system.

6. Electric Filters

An electric filter is a combination of coils and condensers capable of separating electrical waves characterized by a difference in frequency.[1]

The three fundamental forms of filters are commonly termed "low-pass," "high-pass," and "band-pass." A low-pass filter is one which passes currents of frequencies below a particular "cut-off frequency" and attenuates or weakens very greatly currents of higher frequencies. A high-pass filter does the opposite—attenuates below the cut-off frequency and passes above this frequency. A band-pass filter is one which passes currents of frequencies within a definite band fixed by two cut-off frequencies. A low-pass and a high-pass filter connected in series constitute one form of band-pass filter. For any type of filter, the sharpness of cut-off and the amount of attenuation can be controlled at will by suitable design constants.

[1] G. A. Campbell, "Physical Theory of Electrical Wave Filters," *Bell System Tech. Journal*, Nov., 1922.

The stethophone is equipped with five filters whose cut-off frequencies are based on careful analyses of about 100 hospital cases of heart murmurs, râles and breathing sounds. These analyses showed that the sounds of pathological interest to the physician can be grouped into fairly definite frequency regions. When sounds in a particular range of frequencies are of immediate importance, they may be emphasized by suppressing sounds outside of this band.

The frequency characteristics of the filters are shown in Fig. 6.

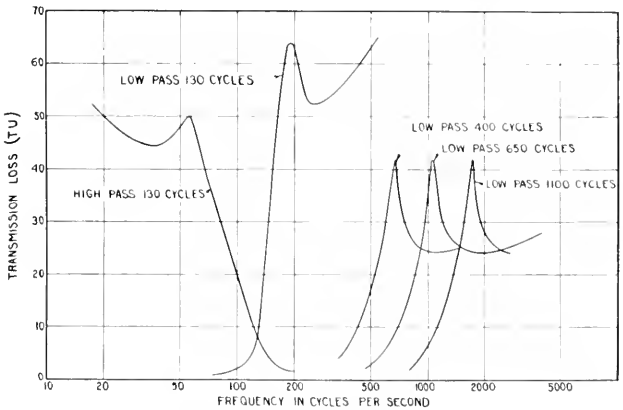


Fig. 6. Loss characteristics of the five filters

For convenience, the cut-off frequency has been defined as that frequency at which the energy is reduced to approximately 1/10 of its original value.

The low-pass filter with a cut-off frequency of 130 cycles is of primary use for reproducing the normal heart sounds and fetal heart sounds in cases where the rate alone is desired. Most of the energy of these sounds is below 100 cycles. With this filter most of the common interfering noises, including the sounds of the human voice, are excluded.

The low-pass 100 cycle filter is particularly useful for observing presystolic and certain low-pitched systolic and diastolic murmurs.

The low-pass 650 cycle filter has been found the most valuable of all five filters. With it, most high-pitched murmurs, low-pitched

râles and certain types of breathing sounds can be observed to the greatest advantage.

The low-pass 1,100 cycle filter passes the higher frequency components of very high-pitched murmurs and high-pitched râles in a majority of cases.

The high-pass 130 cycle filter serves a unique and important purpose. It may be regarded as in value second only to the low-pass 650 cycle filter. In many cases, the loud normal sounds tend to mask or obscure the faint higher-pitched murmurs. The high-pass 130 cycle filter serves to weaken greatly the normal heart sounds so that the murmur sounds occurring in the intervals between the beats appear with its use to be relatively much louder. In this filter, the amount of attenuation in the low frequency region has been made such that the residual low frequency energy and the higher frequency components of the normal heart sounds are just sufficiently audible so that the murmurs may be timed with relation to their positions in the cardiac cycle. This filter is also very useful for weakening the heart sounds when râles or pericardial friction sounds are to be observed in areas where the heart sounds are loud.

The high-pass 130 cycle filter may be connected into the circuit jointly with any one of the low-pass filters, thus making available a group of band-pass filters with a lower cut-off frequency of 130 cycles.

7. Output Receivers

When the stethophone is used for teaching or consultation purposes, a number of high impedance receivers are connected in parallel in the output circuit. Each observer is provided with a single receiver to which the ordinary stethoscope earpieces may be readily

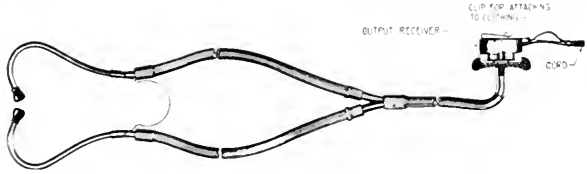


Fig. 7.—The output receiver

attached as shown in Fig. 7. This method of transmitting sounds from the receiver diaphragm to the ears minimizes leakage loss of sound energy and serves effectively to shut out room noises and other annoying sounds. This result could be even better obtained

and with less distortion by providing the receivers with small tips to insert in the ears but at a greater cost for additional receivers. It is perhaps better to use the tubing and earpieces of the ordinary stethoscope as this is the equipment to which physicians are most accustomed and to which the student must accustom himself for future practice. The receiver case is provided with a spring clip for attachment to the clothing. This allows full freedom of both hands for manipulating the transmitter and the control switches of the amplifier, taking notes, etc.

The impedance of the output circuit depends upon the number of receivers in use and, for parallel connection, decreases as the number of receivers is increased. To care for the variable number that may be used at different times, the output transformer has been tapped and a three-way switch provided. By operating this switch, the apparatus can be adjusted to a load varying from 1 to 600 receivers with a maximum transmission loss of 2.5 *TU*.

8. Frequency Characteristics of the Stethophone

The overall frequency characteristics of the stethophone, including the transmitter, the amplifier, and the output receivers, are given in Fig. 8. Two curves are shown. The solid line curve represents

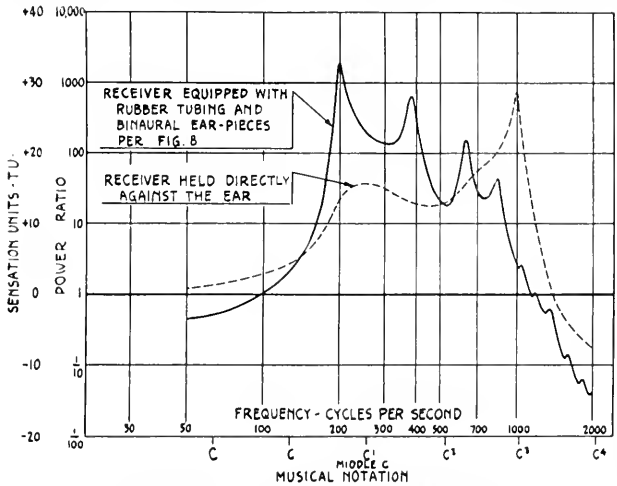


Fig. 8 — Frequency characteristics of the stethophone

operating conditions when the output receivers are equipped with rubber tubing as in Fig. 7, and with the binaural ear-pieces held in the ears. The peaks in this curve are due principally to the resonance in the air columns of the rubber tubing, and correspond to the similar peaks of Fig. 3 for the open bell stethoscope. In order to point out the effect of the stethoscope attachment of the output receiver, a second characteristic is shown by a dotted curve which represents conditions when an output receiver of the same type is held directly against the ear. It is noted that the stethoscope attachment increases the transmission between 150 and 500 cycles per second, and damps the sharp resonant peak of the receiver.

The overall characteristic of the stethophone as employed for auscultation is quite similar to that of the open bell stethoscope. It is desirable that the body sounds as observed by the stethophone should appear the same as in the ordinary stethoscope, particularly in teaching work since the latter is used almost universally in regular practice. If it were deemed desirable for special purposes to avoid the distortion introduced by the stethoscope attachment, receivers with small tips to insert in the ears could be used. For such an arrangement, the overall characteristics could be further improved by using damped receivers which would practically eliminate the sharp peak of the dotted curve in Fig. 8.

9. Installation for Teaching Purposes

When the stethophone is to be used for teaching purposes a permanent wiring or distribution system should be installed with outlets distributed among the seats of the amphitheatre or lecture room.[11] A schematic diagram of such a system is shown in Fig. 9. A distributing pair of feeder wires, preferably shielded, is run between alternate rows of seats below the floor casing, or suitably sheathed to prevent damage. An outlet block "A" of six double contact jacks is mounted on the back of each third seat of alternate rows. Thus, one outlet block will supply six seats, three in front and three in back of the block. Substantial jacks should be used throughout and all receivers should be equipped with rugged plugs. In addition to furnishing jack outlets among the seats, two or three multiple outlet blocks may be installed at the center of the amphitheatre as shown at "B" for the use of guests or others on the floor of the amphitheatre. The output of the stethophone can be connected to the distributing system of the amphitheatre at any one of these outlets. Switch boxes should be installed at various points as at "C" to facil-

itate the localization of an accidental short circuit. If a short circuit should occur in any part of the system this section can thus be disconnected and the balance used independently.

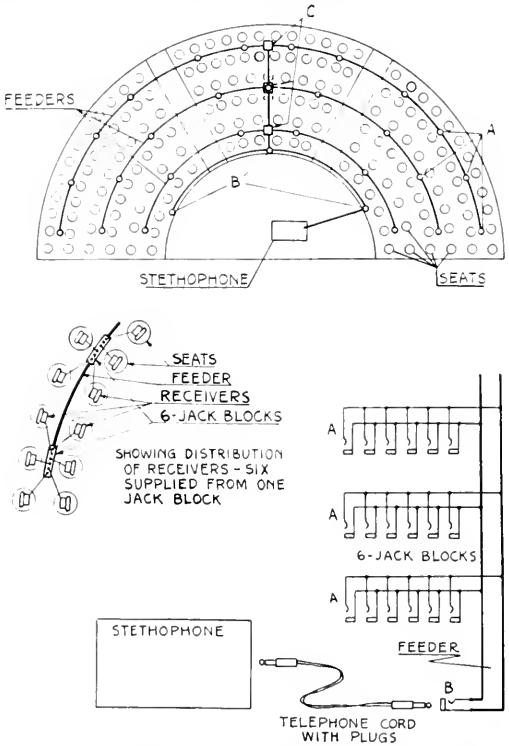


Fig. 9 Wiring installation in an amphitheatre for teaching purposes

In class room lectures, the instructor can make announcements to the students and point out features of particular interest in a convenient and somewhat novel manner without requiring the removal of the stethoscope tubes from their ears. The human body acts as a sounding board for sounds in air that is, when words are spoken

in the vicinity of a patient, the flesh and bone structure vibrates to these sounds. This is particularly true of the areas commonly used in auscultation. The transmitter, resting on the flesh, will pick up these vibrations together with those originating in the body of the patient. The instructor may, therefore, talk to his students by directing his words at that portion of the body to which the transmitter is applied. Best results are obtained with a talking distance of about ten inches. During such announcements, it is essential,

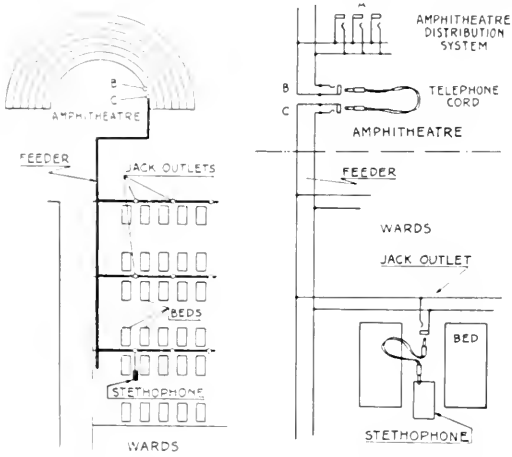


Fig. 10—Wiring installation for the rooms or the wards of a hospital

of course, that the electrical filters be removed from the circuit in order that the important higher frequency components of speech may be transmitted to the receivers. Because of this operating feature, it is obviously necessary to have the patient in a reasonably quiet place.

It is often desirable to reproduce in the lecture hall, the heart and chest sounds of confined patients too ill to be moved. For this purpose, the rooms or the wards of a hospital may be connected by a pair of wires to the lecture room. Such an installation is shown in Fig. 10. Terminal outlets are distributed throughout the rooms or the wards as desired and all are connected to the main feeder wires which communicate with the lecture hall. It is necessary to take the stetho-

phone to the bedside. Long wires from the transmitter to the amplifier cannot be tolerated on account of inductive disturbances from neighboring telephone or other electrical circuits. If desirable, announcements may be made as before by talking close to the body of the patient under observation. In cases where exposure of a patient is inadvisable or where accurate statements pertaining to the seriousness of a disease are preferably withheld from the patient, announcements may be made by talking in a low tone of voice at about one inch distance from the transmitter itself. Reasonably satisfactory reproduction is obtained by this means.

10. Other Applications

Aside from its application to teaching purposes, the stethophone appears to have possibilities in fields which have not yet been thoroughly studied. Further experimental investigation by the medical profession can alone bring out these possibilities.

The possibility of substituting a loud speaker for the individual receivers in the output circuit has been investigated in a preliminary manner. This problem involves certain very fundamental factors relating to the sense of hearing which must be considered carefully. To a remarkable extent, the ear is capable of selective observation. Ordinarily we listen to sounds through a sea of noise to which we become so accustomed that we fail to notice it. However, when listening to sounds near the threshold of audibility, such as the body sounds under consideration, this noise may render the sounds of interest inaudible. In order to hear them, it therefore, becomes necessary to increase the loudness to a point well above that commonly observed by the physician with his stethoscope. This increase in loudness brings within the audible range, sound components ordinarily not heard and changes the quality of the whole as judged by the ear. Such alteration of quality is obviously very unsatisfactory for diagnosis or teaching purposes. Assuming that we had available a perfect loud speaker, one that would transmit the very low and the higher frequency components of faint body sounds without distortion, difficulties would still be presented by the acoustic characteristics of the room in which the loud speaker was placed. All rooms are more or less reverberant. When these sounds are reproduced by a good loud speaker in a small heavily damped sound-proof room, they appear quite natural, but such a room is seldom available practically. None of the ordinary loud speakers with horns will transmit the low frequencies here of interest and would sound very un-

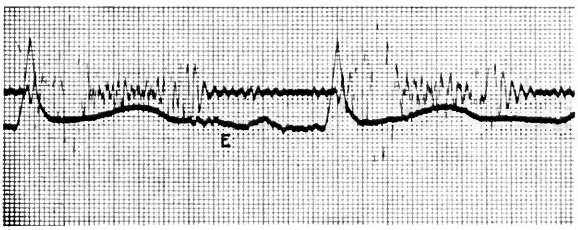
natural even with ideal room conditions. While a loud speaker has been used under proper acoustic conditions to reproduce faint pathological sounds, as murmurs and râles, this does not appear in general to be a practical arrangement. Most arguments, except perhaps that of economy, tend to favor the use of individual output receivers for practically all purposes where critical analysis of sounds is the objective.

Fetal heart sounds as heard through the mother's abdomen are much fainter and require considerably higher amplification than adult heart sounds. Preliminary data indicate that the energy of fetal heart sounds is approximately only 1 50 to 1 500 of the energy of average normal heart sounds. The low pass 130 cycle filter is not only useful for suppressing the extraneous sounds and electrical disturbances which usually attend the use of high amplification, but serves most effectively to eliminate the voice sounds of the patient. At Sloane Hospital in New York City it has been found possible to reproduce clearly on a loud speaker and with a negligible amount of interference, fetal heart sounds which were barely audible in the physician's stethoscope. In these cases no interference was experienced from the maternal heart sounds. However, even in surgical work where the rate of the adult or fetal heart is alone of importance, it is felt that the best plan is to equip an attendant with earpieces attached to a receiver and to make it his chief duty to observe the heart action.

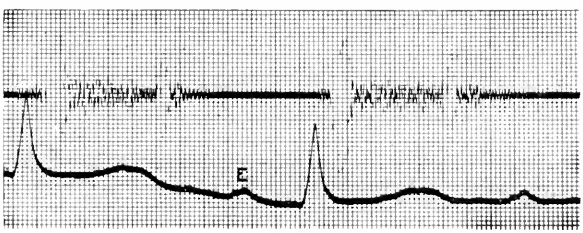
A very important application of the electrical stethoscope is its association with a recording galvanometer for making photographic records of heart and chest sounds. Permanent records of this sort might constitute a valuable addition to the history records of important cases in large hospitals. Some excellent graphical records have already been made.[10] It has been found that such records can be obtained much more easily with the stethophone, principally because of the part played by the electrical filters. The low-pass filter suppresses largely the current fluctuations caused by mechanical vibrations and noise at the apparatus. The high-pass 130-cycle filter is also valuable for bringing out very faint murmurs. By its use, the amplitude of the normal heart sounds can be greatly reduced. When this is done, the amplitude of the faint murmur sounds may be magnified relatively and hence shown very nicely on the record.

This is illustrated in the two charts of Fig. 11, which are presented through the courtesy of Dr. H. B. Williams. The stethophone records are accompanied by simultaneous electro-cardiograms (E) for timing the events. The first record was made with a low pass

650 cycle filter and the second with a band pass 130–1,100 cycle filter and increased amplification. The latter shows the systolic murmur very clearly. With the *LP* 650 filter, the murmur is more or less obscured by very low pitched sounds which may really be a part



Low pass 650 cycle filter



Band pass 130–1100 cycle filter

Fig. 11—Records of a systolic murmur taken with the stethophone and a recording galvanometer

of it, but certainly play a subordinate role in producing the audible sound. The effect of suppressing the low pitched sounds by using the high pass filter is more pronounced in charts of this sort for murmurs which have negligible sound components below 130 cycles per second.

Phonograph records of heart sounds have been made previously.[11] With the stethophone and a special electrical recorder, records of some 15 cases of murmurs and chest sounds have recently been made. The results are very encouraging. All of the characteristics of these sounds, such as the relative intensities of the different components

[11] G. E. Keiper, *Letter J. A. M. A.*, 81:679 Aug. 25, 1923.

and their quality, have been retained remarkably well. The problem of subsequent reproduction of these records has been met satisfactorily in two ways, in both of which the factor of particular concern is the elimination of "needle scratch" noises. First, an electromagnetic reproducer has been used in conjunction with the stethophone. In this case, low-pass filters serve to reduce the scratching noises. Second, the records have been reproduced by attaching the earpieces of the ordinary physician's stethoscope to a special adapter used with a commercial phonograph reproducer. To reduce needle scratch in this case it is only necessary to introduce some form of air passage between the reproducer and the binaural earpieces which acts as a low-pass acoustic filter. The ordinary commercial phonograph is quite unsatisfactory for reproducing these records partly for the same reasons mentioned above relative to the use of loud speakers.

Phonograph records of heart and chest sounds can be employed to some extent for preliminary teaching purposes and do not require much equipment if reproduced acoustically. No patients are required in this case, and the records can be accompanied by the analysis or diagnosis of an expert. It is suggested that phonograph records might be used to advantage as permanent records to follow the progress of disease in important cases.

II. Summary

A summary of the applications and limitations of a new form of electrical stethoscope has been given. However, the extent of its usefulness can be brought out only after it has been placed at the disposal of experienced men in the medical profession. With it, heart murmurs and râles can be magnified and observed with greater clearness than with the ordinary stethoscope. Extremely faint sounds may be heard clearly without great acuity of hearing by inexperienced observers, a thing which has not hitherto been possible. In several instances, murmurs have been discovered with the stethophone which were not discerned initially with the ordinary stethoscope although discernible after having heard them with the more powerful apparatus. In a few of these cases, very faint murmurs, although undoubtedly present could not be heard at all with the ordinary stethoscope. It is felt that the electrical filters have played an important part in such cases. These facts lead us to believe that the stethophone may have real value for diagnosis.

The field of physical research of body sounds has been touched upon but lightly. For special purposes, an endless variety of electrical filters can be used with the stethophone.

Mathematics in Industrial Research[1]

By GEORGE A. CAMPBELL

"Selling" Mathematics to the Industries

THE necessity for mathematics in industry was recognized at least three centuries ago when Bacon said: "For many parts of nature can neither be invented [discovered] with sufficient subtility nor demonstrated with sufficient perspicuity nor accommodated unto use with sufficient dexterity without the aid and intervening of mathematics." Since Bacon's time only a very small part of nature has been "accommodated unto use," yet even this has given us such widely useful devices as the heat engine, the telegraph, the telephone, the radio, the airplane and electric power transmission. It is impossible to conceive that any of these devices could have been developed without "the aid and intervening of mathematics." Present day industry is indeed compelled, in its persistent endeavors to meet recognized commercial needs, to make use of mathematics in all of the three ways pointed out by Bacon. The record of industrial research abundantly confirms his assertion that sufficient subtility in discovery, sufficient perspicuity in demonstration, and sufficient dexterity in use can be achieved only with the aid of mathematics.

There is throughout industry one vitally important common characteristic, uncertainty. In one industry the uncertainty may be due to the supply of raw material, the supply of labor, the supply of brains or the supply of capital. In another industry the uncertainty may be due to the activity of competitors, to fluctuating public demand or to the passage and subsequent interpretations of statutory laws. Still other industries are the playthings of the weather. Whatever the sources of uncertainty it is of vital importance to the industry to reduce to a minimum the hazards due to each of the uncertainties to which it is subjected. To a limited extent hazards may be transferred by means of insurance; but most uncertainties cannot be disposed of in this manner they must be met by the industry individually.

The practice of probabilities, therefore, has a place in every industry. In fact, it occupies the first place in industrial mathematics, barring only the elementary arithmetical operations. It is remarkable how subtle are the mathematical difficulties presented by ap-

[1] Paper read at the International Mathematical Congress, at Toronto, August 11, 1924.

parently innocent problems in the theory of probability. For this reason, mathematicians who are entrusted with the application of probability to industry must have great insight and acumen. Even so, in applying probability to any industry, a beginning should be made with the simpler problems, going on by gradual steps to more and more complicated ones.

Each industry has its own special mathematical problems, which must be considered individually in order to determine where mathematics should be applied. No industrial problem can seem much more hopeless, as a field for exact mathematics, than the subject of electricity as understood in the time of Bacon. It was then a mere collection of curious observations, such as the evanescent attraction of rubbed amber. Persistent observation and careful correlation have, however, brought a large domain of present day electricity under quantitative relations. Electricity is now preeminently a field for mathematics, and all advances in it are primarily through mathematics.

Industrial mathematics will achieve but little unless it is undertaken by persons with suitable aptitudes working under favorable conditions, on problems which have reached the mathematical stage. Industrial mathematical research involves much more than the mechanical application of established mathematical formulas. It involves cooperation in determining the problems to be attacked, in deciding what experimental data are necessary, in obtaining these data, in formulating the mathematical problem, in carrying through the analytical and numerical work, in applying the results to the physical actuality and in practically testing the commercial results achieved. In this cooperation many individuals may be involved and many tentative trials may be necessary in order to determine the solution which best meets all of the commercial conditions.

The cooperation must be effective; it must produce results, and these promptly. Mathematical deductions must be made intelligible and convincing, so that they will eventuate in action even when the indications of theory are apparently contrary to practical experience. This is important because the most valuable theoretical results are often revolutionary.

On the part of the industrial mathematician, powers of observation, clear physical concepts, quick resourcefulness, creative imagination and constant persistency are required. These are rare human qualities. Unless industrial mathematical work is made attractive to men possessing these high talents, the full measure of success cannot be expected. Industrial mathematics must offer a career in

itself, since specialization is required—specialization of a type which eventually disqualifies most men from undertaking other lines of work most effectively.

Mathematics in Electrical Communication

In order to make the foregoing observations somewhat more specific, I will refer to a few applications of mathematics in the industrial research of the Bell Telephone System. This field is selected because I am more familiar with it than with other industrial activities.

Certainty of prediction is the basic requirement in the development and operation of the telephone system; no vital need of the system can be left to chance or to fortuitous development. For this reason, the Bell System is highly organized under research control. The telephone situation is studied as a whole; all departments cooperate; each problem is considered from every point of view. Every attempt is made to master a situation in advance of the necessity of action, so that the most effective and economical means for electrical communication may be adopted with each expansion of the system. Much more than the immediate requirements of the hour must be known; preparation for all eventualities must be made. Fortunately, the executives have carried out this program with a prophetic appreciation of the value and necessity of mathematics.

The importance of the theory and practice of probabilities was recognized as soon as the telephone reached a thoroughly commercial basis. It has proved invaluable during the great expansion which has already carried the number of telephones in the city of New York to over a million. Meeting the peak load demand of the million-odd telephones in New York City, on a practically no-delay basis, with the minimum amount of equipment, is a highly complex and important problem. Without probability studies of the situation, the equipment installed at one point would be inadequate, while at other points it would be superabundant. The superfluous equipment would involve a waste of capital, while the inadequate equipment would mean inconvenience to the public and a loss of possible revenue. Equipment engineering involves a large number of probability problems which are novel, difficult, and financially most important. The aggregate cost of all such studies is large, but the resulting saving to the telephone-using public is much greater. Satisfactory telephone service in metropolitan areas is as dependent upon applied probability as is the success of life insurance.

The telephonic ideal, which is the perfect reproduction of speech, with articulation which is indistinguishable from face-to-face conversation, involves extensive and exhaustive investigations in many fields, in particular in mechanics, acoustics and electromagnetism, since each telephonic conversation involves oscillations in the air, in solids and in the ether. Fortunately, the foundations of the mathematical theory in these three fields had been securely laid by the time Alexander Graham Bell effected their harmonious cooperation in his first telephone. It is impossible for us to be too well informed concerning the consequences of the mathematical laws in these three fields.

It is characteristic of many problems encountered in industry that a great number of independent variables are involved, far too great a number for the best solution to be reached simply by trained judgment. Consider the transposition problem of the telephone system, which is this: on pole lines, long lines between cities, for example, several wires sometimes a great number of wires are strung along in close proximity. Each pair of wires receives inductive effects from the electric waves carried by every other pair, producing so-called crosstalk. To reduce such effects, the pairs of wires are transposed according to a set plan; that is, the positions of the two wires are interchanged, an expedient analogous to the twisting of a pair of wires. It is necessary to consider not only the ideal location of the transpositions in each pair of wires, but also the practical irregularities which occur in the actual placing of the transpositions. One of the practical problems, in fact, is to determine the allowable tolerances limiting the irregularities in the location of loading coils and transpositions, since these irregularities modify the crosstalk and also the transmission efficiency by an amount which must be determined by the laws of probability.

Transpositions were originally introduced with complete success about thirty years ago, and yet at the present time this subject is being more actively studied than ever; this is due to the extended use of phantom circuits and the new uses of carrier frequencies, that is, high-frequency speech-carrying currents which are superposed on ordinary telephony.

To illustrate the way in which problems in industrial mathematics become, step by step, more complex by the progressive inclusion of one factor after another, brief reference may be made to the loaded cable circuit. The first successful telephone cable circuits could be treated mathematically on the basis of Kelvin's simple cable diffusion theory. To allow for the ignored inductance and to deter-

mine the effect of added inductance, Heaviside's much more complete transmission formulas were employed somewhat later. The next stage was to allow for the effect of inductance which was not uniformly distributed, but lumped at regular intervals. Here the steady state solution for sinusoidal vibrations of a loaded string was employed, and the cutoff frequency due to internal reflections at the loading coils determined. But with loaded cables of great length, extending from New York to Chicago and beyond, the transient state may be of such duration as to require consideration. The loaded line does not transmit the impulse as a whole, but breaks it up by reflection and transmission at each loading coil. Therefore some of the impulses arrive after a few short backward reflections, while other impulses may travel many times the length of the line, due to reflections back and forth at many of the thousand loading coils in the circuit. The calculation of the transient state at the receiving end, due to the arrival of these impulses in groups, one after another, involved the calculation of Bessel functions up to order 2000 and subsequent integration by an application of the principle of stationary phase to Fourier's integral.

Industrial Mathematics as a Career

It is true that the mathematician who takes up industrial work is not entirely free to set his own problems; the industry which he has chosen provides these and it demands concentration upon them. Such problems are often less inviting than the clear-cut, tractable problem which the pure mathematician is at liberty to set himself. Industrial problems may be most complicated to frame and they may admit only of approximate solution by laborious numerical methods. In addition to delimiting the nature of his problems, the imperative needs of industry set time limits for their solution, and the nature of industry demands a financial profit from industrial mathematics. But these restrictions of industry should not make the work less attractive. On the contrary, restrictions disclose the master. There is an inspiration in overcoming even the humblest difficulty standing in the path of progress. Restrictions, even in the case of the most gifted, may be beneficial in concentrating activities, thereby making up in depth what may seem lacking in breadth.

The industrial mathematician may have a chance to attack many large-scale investigations which would be impossible, except under the patronage of industry, because of the exceptional material equip-

ment and widely sustained cooperation required. Some of the opportunities offered by cheap electrical power from Niagara, by high-voltage electric power lines, and by large steam turbines may be mentioned. It is often left to the industrial mathematician to reap the harvest from seed sown under adverse circumstances by pure mathematicians.

The industrial mathematician may hope to make some return for the debt which he owes the pure mathematician. He may introduce new mathematical problems, of which industry is an inexhaustible source. He may point out the application of pure mathematical results, stimulating further investigations along the same lines. He may assist mathematicians generally by promoting the preparation of needed tables and by creating a commercial demand for calculating machines and other brain-saving devices.

The opportunities presented by industrial mathematics are boundless, because mathematics is the key to extrapolation in time, and industry is absolutely dependent upon prediction. The position of mathematicians in industry must eventually correspond with the importance of the function which they may perform.

Training for Industrial Mathematics

In industry we are concerned with mathematics not as an objective, but only as a tool. It follows that the required training in mathematics should develop a wide acquaintance with the available mathematical tools and practical skill in their use. It is important to note the distinction between the using of tools and the making of tools. Under primitive conditions the workman makes his own tools, but in a highly organized society the tools are made by specialists, who provide the workman with an endless variety of implements superior to anything which he himself could make. By long experience the tool designer has discovered how best to adapt the tool to its intended use in order to economize the workman's time and energy as much as possible. Furthermore, the substitution of one tool for another with the minimum number of motions is made possible by the use of interchangeable parts and systematically arranged cabinets.

But no complete line of mathematical tools is for sale across the counter; only a limited number of numerical and algebraic tables and a few types of calculating machines are supplied as ready-made tools. By far the larger part of known mathematical tools must be sought for in the literature of the subject, but there they may be

difficult to find and isolate in the form best adapted for the purpose in hand. What is very greatly needed at the present time is a compendium or unabridged dictionary of mathematical results concisely and uniformly stated, and systematically classified for convenient reference. What I have in mind is not a mere handbook of applied mathematics, but a statement of theorems and formulas and tabulated results expressed in the language of pure mathematics, and comparable in scope and size with the "Encyklopädie der Mathematischen Wissenschaften." Preparation of such a compendium would be a tremendous undertaking, but it would also be of the greatest value. To such a collection of tools the industrial mathematician would turn for the appropriate tool as each new problem arises.

I would have the university training of the industrial mathematician based upon such a compendium by means of judicious sampling, at many points, under competent leadership. He would thus become familiar with his source book as a whole and thereafter turn to it instinctively and use it with confidence. At the present time, when the average text-book is held in low esteem and nothing has been substituted which adequately fills the gap, the student of mathematics leaves the university with a five-foot shelf of notebooks, together with what he carries in his head. Neither the memory nor the notebook is likely to be a reliable source of information when a particular result is needed for the first time, ten years later. It then becomes necessary for him to take the time to deduce the result from first principles, or to hunt up lecture notes, a text-book or original paper and waste much valuable time picking up the thread of the argument. The sampling to which I have referred should not be that of a dilletante; it should be an intensive grounding in the fundamental concepts and methods of mathematics, and the development *ab initio* of several well distributed branches of mathematics.

The combination of mathematical ability with an observant mind is as desirable as it is rare. The university training should include non-mathematical courses adapted for developing the powers of observation, or at least an appreciation of the necessity of cooperating with others who are observant. A study of the natural sciences, accompanied by experimental work, should be of great value. It is of course, difficult to be reasonable and not ask the impossible of the university in the training of any specialist. We recognize that, at best, only a beginning can be made at the university, but this beginning should include the fundamentals and should not attempt

to impart details of current industrial practice. These details are best acquired in the industrial environment itself. Self-training in fundamentals, on the other hand, is much more difficult, and is not likely to go far, unless a start has been made under the favorable conditions afforded by the university.

What I have tried to emphasize is that industry can realize its greatest possibilities only with the aid of mathematicians, and that mathematicians can find opportunities in industry worthy of their powers, however great those powers may be. To ensure the success of industrial mathematics the industry must inaugurate mathematical research as early as possible, so that ample time may be afforded for the gradual accumulation of information upon which mathematics may be securely based, and for deriving quantitative results before the necessity for commercial action arrives. The industrialist must also be ready to give the mathematician's conclusions a sympathetic trial even though they run contrary to established precedent. Above all, industry needs mathematicians of an especially broad type—men whose interests naturally extend beyond their special field, and who are flexible enough to cooperate with non-mathematicians. These industrial mathematicians must inspire confidence by their firm grasp of physical realities, by the relevance of their mathematics, and by the ability to present their results clearly and convincingly.

The Building-up of Sinusoidal Currents in Long Periodically Loaded Lines

By JOHN R. CARSON

IMPORTANT information regarding the excellence of a signal transmission system is deducible from a knowledge of the mode in which sinusoidal currents "build-up" in response to suddenly applied sinusoidal electromotive forces, since on the character and duration of the "building-up" process depend the speed and fidelity with which the circuit transmits rapid signal fluctuations.[1] The object of this note is to disclose and discuss general formulas and curves which describe the building-up phenomena, as a function of the line characteristics and the frequency of the applied e.m.f., in the extremely important case of long periodically loaded lines. The formulas in question are approximate but give accurate engineering information and are applicable to all types of periodic loading under two restrictions: (1) the line must be fairly long, that is, comprise at least 100 loading sections, and (2) it must be approximately equalized, as regards *absolute* steady-state values of the received current, in the neighborhood of the applied frequency. Fortunately these conditions are usually satisfied in practice in those cases where the building-up phenomena are of practical engineering importance. Furthermore, the formulas to be discussed supply a means for the accurate and rapid comparison of different types of loading in correctly engineered lines.

The building-up process may be precisely defined and formulated as follows: Suppose that an e.m.f., $E \cos \omega t$, is suddenly applied, at reference time $t = o$, to a network of transfer impedance

$$Z(i\omega) = |Z(i\omega)| \cdot \exp [iB(\omega)]. \tag{1}$$

The resultant current, $I(t)$, may be written as

$$I(t) = \frac{1}{2} \frac{E}{|Z(i\omega)|} \{(1+\rho) \cos [\omega t - B(\omega)] + \sigma \sin [\omega t - B(\omega)]\}, \tag{2}$$

$$= \frac{1}{2} \sqrt{(1+\rho)^2 + \sigma^2} \frac{E}{|Z(i\omega)|} \cos [\omega t - B(\omega) + \theta], \tag{3}$$

where

$$\theta = \tan^{-1}(\sigma \div \rho).$$

Evidently the functions ρ and σ must be -1 and o respectively for negative values of t, and approach the limits $+1$ and o as $t = \infty$.

[1] For published discussions of the "building-up" of sinusoidal currents in loaded lines, see Clark, *Journ. A.I.E.E.*, Jan., 1923, Kupfmuller, Telegraphen u. Fernsprech-Technik, Nov., 1923; Carson, Trans. A.I.E.E., 1919.

In an engineering study of the building-up process we are principally concerned with the *envelope* of the oscillations, which, by (3), is proportional to

$$\frac{1}{2}\sqrt{(1+\rho)^2+\sigma^2}.$$

The problem is therefore to determine the functions ρ and σ and to examine the effect of the applied frequency $\omega = 2\pi$ and the characteristics of the circuit on their rate of building-up and mode of approach to their ultimate steady values.

Two propositions will now be stated which cover the building-up process in the practically important cases. Since the line is assumed to be approximately equalized, as regards the absolute value of the received current in the neighborhood of the applied frequency $\omega = 2\pi$, the building-up process depends only on the total phase angle $B(\omega)$. The successive derivatives of the phase angle with respect to ω will be denoted by $B'(\omega)$, $B''(\omega)$, $B'''(\omega)$, $B^{\text{IV}}(\omega)$, etc.

Case I. $B''(\omega) = 0$ and $\sqrt{B''(\omega)}$ 2! large compared with $\sqrt[3]{B'''(\omega)}$ 3!.

The envelope of the oscillations in response to an e.m.f. $E \cos \omega t$ *applied at time* $t = o$, *is proportional to*

$$\frac{1}{2}\sqrt{(1+\rho)^2+\sigma^2} \tag{4}$$

where

$$\rho = C(x^2) + S(x^2), \tag{5}$$

$$\sigma = C(x^2) - S(x^2), \tag{6}$$

$$x = \frac{t - B'(\omega)}{\sqrt{2B''(\omega)}} = \frac{t'}{\sqrt{2B''(\omega)}}, \tag{7}$$

and $C(x)$, $S(x)$ *are Fressel's Integrals to argument* x.

The envelope therefore reaches 50 per cent. *of its ultimate steady value at time* $t = \tau = B'(\omega)$ *and its rate of building-up is inversely proportional to* $\sqrt{B''(\omega)}$.

The curve of Fig. 1 is a plot of the envelope function $\frac{1}{2}\sqrt{(1+\rho)^2+\sigma^2}$ to the argument x and is therefore applicable to all types of loading and lengths of line, subject to the restrictions noted above.

Case II. $B''(\omega) = 0$; $B'''(\omega) \neq 0$ and $\sqrt[3]{B'''(\omega)}$ 3! large compared with $\sqrt[4]{B^{\text{IV}}(\omega)}$ 4!.

The envelope of the oscillations is proportional to

$$\frac{1}{3}+\frac{1}{2}\int_0^y A(\mu)d\mu \tag{8}$$

where $A(\mu)$ *is Airey's Integral*[2] *and*

$$y=\left(\frac{2}{\pi}\right)^{2/3}\frac{t-B'(\omega)}{\sqrt[3]{B'''(\omega)}\ 3!} \tag{9}$$

$$=t'\sqrt[3]{\frac{24}{\pi^2 B'''(\omega)}}. \tag{10}$$

At time $t=B'(\omega)$ *the envelope* N *has reached* $\frac{1}{3}$ *of its ultimate steady value and its rate of building-up is* inversely proportional to $\sqrt[3]{B'''(\omega)}$.

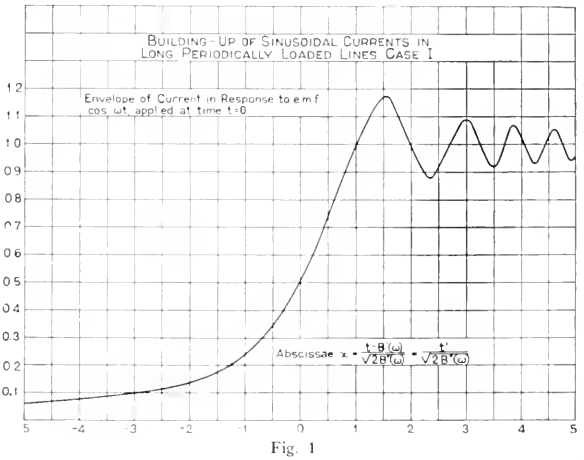


Fig. 1

The curve of Fig. 2 is a plot of the envelope function $\frac{1}{3}+\frac{1}{2}\int_0^y A(\mu)d\mu$ to the argument y and is therefore of general applicability under the circumstances where case II obtains.

The practical value of the foregoing propositions resides in the fact that they enable us to calculate two important criteria of the transmission properties of the line: (1) the variation with respect to frequency of the time interval τ required for the current to build-up to

[2] See Watson, Theory of Bessel Functions, p. 190.

its proximate steady-state value; and (2) its rate of building-up at time $t = \tau$.

As will be seen in connection with the proof given below, the formulas of the foregoing propositions are approximate. Provided, however, that the lines to which they are applied are long and provided that the

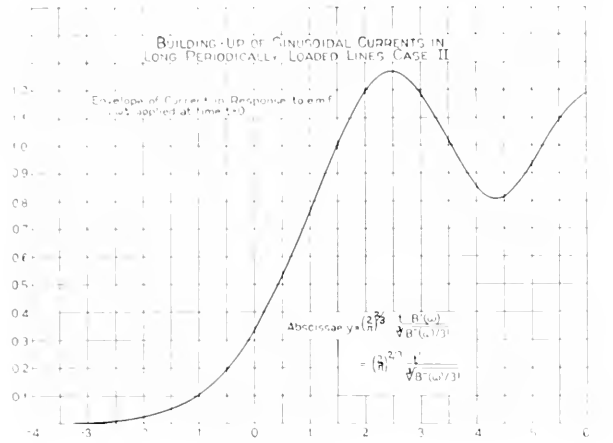


Fig. 2

applied frequency is such that the restrictions underlying either I or II are satisfied, their accuracy is quite sufficient for engineering purposes, such as the design of loading systems, or a study of the comparative merit of different types of loading.

Before proceeding with the mathematical proof, the formulas will be applied to the interesting and important case of an ideal non-dissipative periodically coil-loaded line of N sections in length and cut-off frequency $\omega = 2\pi$. For this line it is easy to show that [3]

$$B'(\omega) = \frac{2N}{\omega_c}\frac{1}{\sqrt{1-\bar{\omega}^2}} = N\beta'(\omega),$$

$$B''(\omega) = \frac{2N}{\omega_c^2}\frac{\bar{\omega}}{(1-\bar{\omega}^2)^{3/2}} = N\beta''(\omega),$$

$$B'''(\omega) = \frac{2N}{\omega_c^3}\frac{1+2\bar{\omega}^2}{(1-\bar{\omega}^2)^{5/2}} = N\beta'''(\omega),$$

[3] The following formulas assume that the line is closer to its characteristic impedance. $\beta\,\omega$ is then the phase angle per loading section of the line.

where w denotes ω/ω_c. It follows that

$$t' = t - \frac{2N}{\omega_c}\left(1 - \sqrt{1-w^2}\right)$$

and that the oscillations build-up to the proximate steady-state in a time interval[4] $\tau = 2N/\omega_c\sqrt{1-w^2}$ after the voltage is applied.

Case I, it will be observed, does not hold for $\omega = 0$ since $B''(0) = 0$. The condition that Case I shall apply is that

$$\sqrt[6]{18N} \cdot \frac{(1-w^2)^{1/12}}{(1+2w^2)^{1/3}}$$

shall be substantially greater than unity. Hence Case I applies only when $1/\sqrt[3]{18N} < w < 1$. This however, includes the important part of the signalling frequency range in properly designed lines, provided that they are long ($N \geq 100$).

In the range of applied frequencies, therefore, corresponding to $1/\sqrt[3]{18N} < w < 1$, the current reaches 50 per cent. of its ultimate steady value in a time interval $\frac{2N}{\omega_c}\left(1 - \sqrt{1-w^2}\right)$ after the voltage is applied and its rate of building-up at this time is proportional to

$$\frac{\omega_c}{4N}\sqrt{(1-w^2)^{3/4}}.$$

For the non-dissipative coil-loaded line $B''(\omega) = 0$ when $\omega = 0$, and Case II applies. Consequently when $\omega = 0$, the oscillations reach 1/3 of the ultimate steady value at time $t = 2N/\omega_c$, at which time their rate of building-up is proportional to

$$\omega_c \sqrt[3]{\frac{12}{\pi^2 N}}$$

The foregoing formulas have been shown to be in good agreement with experimental results, and have been applied to the design of loaded lines in the Bell System.

Mathematical Discussion

The functions ρ and σ of equations (2) and (3) can be formulated as the Fourier integrals

[4] It will be noted that this formula breaks down at $\omega = \omega_c$ or $w = 1$.

$$\rho = \frac{1}{\pi}\int_0^\infty \frac{d\lambda}{\lambda} \sin t\lambda \,[P_\omega(\lambda) + P_\omega(-\lambda)]$$

$$- \frac{1}{\pi}\int_0^\infty \frac{d\lambda}{\lambda} \cos t\lambda \,[Q_\omega(\lambda) - Q_\omega(-\lambda)], \tag{11}$$

$$\sigma = \frac{1}{\pi}\int_0^\infty \frac{d\lambda}{\lambda} \sin t\lambda \,[Q_\omega(\lambda) + Q_\omega(-\lambda)]$$

$$+ \frac{1}{\pi}\int_0^\infty \frac{d\lambda}{\lambda} \cos t\lambda \,[P_\omega(\lambda) - P_\omega(-\lambda)], \tag{12}$$

where

$$P_\omega(\lambda) = \frac{A(\omega+\lambda)}{A(\omega)} \cos\,[B(\omega+\lambda) - B(\omega)], \tag{13}$$

$$Q_\omega(\lambda) = \frac{A(\omega+\lambda)}{A(\omega)} \sin\,[B(\omega+\lambda) - B(\omega)], \tag{14}$$

and $A(\omega) = 1 - Z(i\omega)$.

These formulas are directly deducible from the fact that the applied e.m.f., defined as zero for negative values of t and $E\cos\omega t$ for $t \geq 0$, can itself be expressed as

$$\frac{E}{2}\cos\omega t\left[1 + \frac{2}{\pi}\int_0^\infty \frac{d\lambda}{\lambda} \sin t\lambda\right].$$

In the practically important case where $B'(\omega)$ is finite, it is of advantage to introduce the transformation $t' = t - B'(\omega)$, and to write:

$$\rho = \frac{1}{\pi}\int_0^\infty \frac{d\lambda}{\lambda} \sin t'\lambda \,[U_\omega(\lambda) + U_\omega(-\lambda)]$$

$$- \frac{1}{\pi}\int_0^\infty \frac{d\lambda}{\lambda} \cos t'\lambda \,[V_\omega(\lambda) - V_\omega(-\lambda)], \tag{15}$$

$$\sigma = \frac{1}{\pi}\int_0^\infty \frac{d\lambda}{\lambda} \sin t'\lambda \,[V_\omega(\lambda) + V_\omega(-\lambda)]$$

$$+ \frac{1}{\pi}\int_0^\infty \frac{d\lambda}{\lambda} \cos t'\lambda \,[U_\omega(\lambda) - U_\omega(-\lambda)], \tag{16}$$

where

$$U_\omega(\lambda) = \frac{A(\omega+\lambda)}{A(\omega)} \cos\,[B(\omega+\lambda) - B(\omega) - \lambda B'(\omega)], \tag{17}$$

$$V_\omega(\lambda) = \frac{A(\omega+\lambda)}{A(\omega)} \sin\,[B(\omega+\lambda) - B(\omega) - \lambda B'(\omega)]. \tag{18}$$

The foregoing formulas for ρ and σ are exact subject to certain restrictions on the impedance function $Z(i\omega)$ which are satisfied in the case of periodically loaded lines. Their useful application to the problem under consideration depends, however, on the following approximations.

First it will be assumed that the line is approximately equalized, as regards *absolute* value of steady state received currents in the neighborhood of the impressed frequency ω 2π. By virtue of this assumption, which is more or less closely realized in practice, the ratio $A(\omega+\lambda)$ $A(\omega)$ may be replaced by unity in the integrals (15) and (16), and in equations (17) and (18). It is further assumed that the function

$$B(\omega+\lambda)-B(\omega)-\lambda B'(\omega)$$

admits of power series expansion, so that

$$U_\omega(\lambda)=\cos\left[(h_2\lambda)^2+(h_3\lambda)^3+\ .\ .\right], \tag{19}$$

$$V_\omega(\lambda)=\sin\left[(h_2\lambda)^2+(h_3\lambda)^3+\ .\ .\right], \tag{20}$$

where

$$h_n=\frac{1}{n!}\frac{d^n}{d\omega^n}B(\omega)=\frac{1}{n!}B^{(n)}(\omega).$$

By virtue of the foregoing ρ and σ are given by

$$\rho=\frac{2}{\pi}\int_0^\infty\frac{d\lambda}{\lambda}\sin\left[t'\lambda-(h_3\lambda)^3-(h_5\lambda)^5\ .\ .\right]\cdot\cos\left[(h_2\lambda)^2+(h_4\lambda)^4+\ .\ .\right], \tag{21}$$

$$\sigma=\frac{2}{\pi}\int_0^\infty\frac{d\lambda}{\lambda}\sin\left[t'\lambda-(h_3\lambda)^3-(h_5\lambda)^5\ .\ .\right]\cdot\sin\left[(h_2\lambda)^2+(h_4\lambda)^4+\ .\ .\right]. \tag{22}$$

Now if the line is very long the integrals (11) and (12) may be replaced by the approximations

$$\rho=\frac{2}{\pi}\int_0^\infty\frac{d\lambda}{\lambda}\sin\left[t'\lambda-(h_3\lambda)^3\right]\cdot\cos\left(h_2\lambda\right)^2, \tag{23}$$

$$\sigma=\frac{2}{\pi}\int_0^\infty\frac{d\lambda}{\lambda}\sin\left[t'\lambda-(h_3\lambda)^3\right]\cdot\sin\left(h_2\lambda\right)^2. \tag{24}$$

In other words we retain only the leading terms in the expansion of the function

$$B(\omega+\lambda)-B(\omega)-\lambda B'(\omega).$$

The justification for this procedure depends on arguments similar to those underlying the Principle of Stationary Phase (see Watson, Theory of Bessel Functions, p. 229). Furthermore the upper limit

∞ may be retained without serious error, even when the line cuts off at a frequency $\omega = 2\pi$, provided the line is sufficiently long, and the frequency $\omega = 2\pi$ not too close to the cut-off frequency $\omega_c = 2\pi$.

The formal solutions of the infinite integrals (23) and (24) can be written down by virtue of the following known relations:

$$\frac{2}{\pi}\int_0^\infty \frac{d\lambda}{\lambda}\sin t'\lambda \cdot \cos(h_2\lambda)^2 = C(x^2) + S(x^2), \tag{25}$$

$$\frac{2}{\pi}\int_0^\infty \frac{d\lambda}{\lambda}\sin t'\lambda \cdot \sin(h_2\lambda)^2 = C(x^2) - S(x^2), \tag{26}$$

where $C(x^2)$ and $S(x^2)$ are Fresnel's Integrals to argument x^2, and $x = t'/2h_2$.

$$\frac{2}{\pi}\int_0^\infty \frac{d\lambda}{\lambda}\sin\left[t'\lambda - (h_3\lambda)^3\right] = -\frac{1}{3} + \int_0^y A(y)dy \tag{27}$$

where $A(y)$ denotes Airey's Integral (see Watson, Theory of Bessel Functions) and $y = (2/\pi)^{2/3}(t'/h_3)$.

By aid of the preceding,

$$\rho = \left\{1 + \frac{\mu^3}{1!}\frac{d^3}{dx^3} + \frac{\mu^6}{2!}\frac{d^6}{dx^6} + \cdots + \cdots\right\}\cdot\left\{C(x^2) + S(x^2)\right\}, \tag{28}$$

$$\sigma = \left\{1 + \frac{\mu^3}{1!}\frac{d^3}{dx^3} + \frac{\mu^6}{2!}\frac{d^6}{dx^6} + \cdots + \cdots\right\}\cdot\left\{C(x^2) - S(x^2)\right\}, \tag{29}$$

where $\mu = (h_3/2h_2)$.

This is the appropriate form of solution when (h_3/h_2) is less than unity.

On the other hand when (h_3/h_2) is greater than unity, the appropriate form of solution is

$$\rho = \left\{1 - \frac{\nu^4}{2!}\frac{d^4}{dy^4} + \frac{\nu^8}{4!}\frac{d^8}{dy^8} + \cdots\right\}\cdot\left\{-\frac{1}{3} + \int_0^y A(y)dy\right\}, \tag{30}$$

$$\sigma = \left\{\nu^2\frac{d^2}{1!\,dy^2} - \frac{\nu^6}{3!}\frac{d^6}{dy^6} + \cdots\right\}\cdot\left\{-\frac{1}{3} + \int_0^y A(y)dy\right\}, \tag{31}$$

where $\nu = \left(\frac{2}{\pi}\right)^{\frac{2}{3}}\left(\frac{h_1}{h_2}\right)$.

While no thorough investigation has been made, it appears probable that for all values of the ratio h_3/h_2, either (28), (29) or (30), (31) will be convergent. However, in practice it is sufficient for present

purposes to deal only with the cases where h_3/h_2 is either small or large compared with unity, and to use the following approximations:

(1) (h_3/h_2) *small compared with unity.*

$$\rho = C(x^2) + S(x^2),$$
$$\sigma = C(x^2) - S(x^2),$$
$$x = (t'/2h_2)^2,$$
$$t' = t - B'(\omega).$$

(2) (h_3/h_2) *large compared with unity.*

$$\rho = -\frac{1}{3} + \int_0^y A(y)dy,$$
$$\sigma = 0,$$
$$y = (2/\pi)^{2/3}(t'/h_3).$$

Transmission Characteristics of Electric Wave-Filters

By OTTO J. ZOBEL

SYNOPSIS—The transmission loss characteristic of a transmitting network as a function of frequency is an index of the network's steady-state selective properties. Methods of calculation heretofore employed to determine these characteristics for composite wave-filters are long and tedious. This paper gives a method for such determinations which greatly simplifies and shortens the calculations by the introduction of a system of charts. Account is taken of the effects of both wave-filter dissipation and terminal conditions. The method is based upon formulae containing new parameters, called "image parameters," which are the natural ones to use with composite wave filters.

A detailed illustration of the use of this chart calculation method is given and the transmission losses so obtained are found to agree, except for differences which in practice are negligible, with those obtained by long direct computation.

In the Appendix are derived two sets of corresponding formulae which are applicable to a linear transducer of the most general type, namely, an active, dissymmetrical one; the one set contains image parameters and the other set recurrent parameters. An impedance relation is found to exist between the four open-circuit and short-circuit impedances of a linear transducer even in the most general case. Reduction of these formulae to the more usual case of a passive linear transducer is also made, those containing the image parameters being especially applicable to the case of composite wave-filters.

I. INTRODUCTION

ELECTRIC wave-filter characteristics and systematic methods of deriving them have been considered in previous numbers of this Journal.[1] This paper deals with a simple and rapid method of calculating the steady-state transmission losses of wave-filter networks over both the transmitting and attenuating frequency bands, including the effects of dissipation and wave-filter terminal conditions. Such transmission loss determinations are essential in showing the selective characteristics of these networks and serve as important guides in meeting given design requirements.

General formulae for any dissymmetrical linear transducer are derived in terms of new parameters, called image parameters. One of the formulae is fundamental to the solution of the present problem and is particularly well adapted to calculations in composite wave-filter structures. These parameters of such a composite structure, being readily obtainable from those of its parts, are the natural parameters to use in this case. The formula possesses, among others,

[1] Physical Theory of the Electric Wave-Filter, G. A. Campbell, B. S. T. J., Nov., 1922; Theory and Design of Uniform and Composite Electric Wave-Filters, O. J. Zobel, B. S. T. J., Jan., 1923. Transient Oscillations in Electric Wave Filters, J. R. Carson and O. J. Zobel, B. S. T. J., July, 1923.

the advantage over other formulae and calculation methods of requiring for every alteration in a composite wave-filter only a partial recalculation rather than a more or less complete one. In addition, by its use much of the otherwise necessary calculation can be eliminated through the means of graphical representation.

The main object of this paper is to present this chart calculation method of determining composite wave-filter transmission losses, giving its theory, the necessary charts, and an application of its use.

Structure of Wave-Filter Networks

The ladder type of recurrent network having physical series and shunt impedances z_1 and z_2, respectively, as shown in Fig. 1, is the

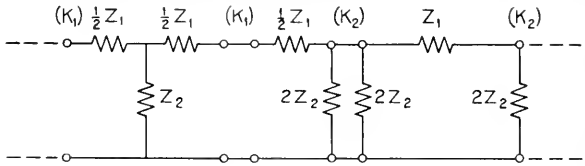


Fig. 1—Ladder Type Recurrent Network

one most frequently employed for wave-filters. Also any passive transducer having two pairs of terminals can theoretically be reduced to the form of the ladder type. Hence, in what follows the ladder type terminology will be used understanding, however, that other structural types may also be included; for example, such as are derivable from the ladder type by the substitution of an equivalent transformer with mutual impedance for T or Π connected inductances, or the lattice type. The figure illustrates, from left to right, one mid-series section, one mid-half section (a dissymmetrical half section terminated at mid-series and mid-shunt points), and one mid-shunt section all connected so as to give a uniform structure.[2] The characteristic impedances of the ladder type at mid-series and at mid-shunt points are K_1 and K_2, respectively.

The majority of wave-filter networks are not uniform throughout their length but have a composite structure designed as given in the paper (B. S. T. J., Jan., 1923) already mentioned. That is, the interior or *mid-part* of a composite wave-filter consists of mid-series,

[2] The same network may also be considered as made up in other ways; for example, two mid-series and one mid-half sections, one mid-half and two mid-shunt sections, or five mid-half sections.

mid-shunt, and mid-half sections, usually dissimilar, so connected serially and of such types that at any junction the terminations of the two adjacent types correspond to an equivalent image impedance. The use of dissimilar sections gives a resultant selective characteristic different from that possible with a uniform type. At the terminals of the network there need not be complete full or half sections; this

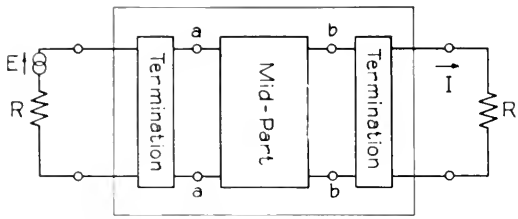


Fig. 2—General Composite Wave-Filter Network

is represented in Fig. 2 by the wave-filter parts external to the *mid-part* which latter is included between terminals *a* and *b*.

The terminations of wave-filter networks specifically considered in detail here include all terminations which have been found to be practical. In any particular class of wave-filter they are all closely related to the "constant k" wave-filter ($z_{1k}z_{2k}=k^2=$constant) of that class and are of the following four types,[3] being designated by their characteristic impedances in corresponding ladder type structures.

[3] It is assumed that the reader is familiar with the terms and notation used in the paper, B.S.T.J., Jan., 1923.

If z_{1k} and z_{2k} are the series and shunt impedances of the "constant k" wave-filter, the corresponding series and shunt impedances of the mid-series "constant k" equivalent M-type are expressible as

$$z_{11}=mz_{1k},$$

and

$$z_{21}=\frac{1-m^2}{4m}z_{1k}+\frac{1}{m}z_{2k};$$

and of the mid-shunt "constant k" equivalent M-type

$$z_{12}=\frac{1}{\dfrac{1}{mz_{1k}}+\dfrac{1}{\dfrac{4m}{1-m^2}z_{2k}}},$$

and

$$z_{22}=\frac{1}{m}z_{2k}.$$

Here the condition $0<m\leqq 1$ is sufficient for a physical structure in all cases.

1, mid-shunt of a mid-series "constant k" equivalent M-type, $(K_{21}(m))$;
2, mid-series of a mid-shunt "constant k" equivalent M-type, $(K_{12}(m))$;
3, x-shunt of the "constant k" wave-filter, (K_{x2}); and
4, x-series of the "constant k" wave-filter, (K_{x1}).

The terminations in $K_{21}(m)$ and $K_{12}(m)$ are employed, as stated in a previous paper,[4] when it is desirable to obtain certain selective characteristics and to minimize reflection losses at the important frequencies to be transmitted, a minimum for the latter occurring where $m = .6$ approximately. The x-shunt "constant k" termination, designated by the characteristic impedance K_{x2}, is a "constant k" type termination in a shunt element, whose admittance is x times (x from 0 to 1) that of a full shunt "constant k" admittance, $\frac{1}{z_{2k}}$; that is, a shunt element whose impedance is $\frac{z_{2k}}{x}$. Similarly, the x-series "constant k" type termination corresponding to the characteristic impedance K_{x1} ends in a series element of impedance xz_{1k}. In the usual case where two or more wave-filter networks having different transmitting bands are associated together, either termination 1 or 2 is suitable for the unconnected terminals, while terminations 3 and 4 are adapted to the terminals connected in series or in parallel, respectively. For two complementary wave-filters, thus connected, minimum reflection losses occur at their junction with a transmission line if $x = .8$ approximately. A relation between this case and termination $K_{12}(m)$ and $K_{21}(m)$ has previously been pointed out, namely, that the series or parallel connected wave-filters have a combined impedance in the transmitting band of either wave-filter approximately like that of $K_{12}(m)$ or $K_{21}(m)$, respectively.

Where the termination is x-shunt or x-series we shall consider that the *mid-part* of the wave-filter begins at the mid-shunt or the mid-series point, respectively, irrespective of whether x is greater or less than .5. Also the *mid-part* need not here necessarily begin in the "constant k" type, but in any wave-filter having an equivalent characteristic impedance.

Transmission Loss

In the design of a wave-filter network the magnitude of k for the corresponding "constant k" wave-filter has been taken equal to the

[4] B. S. T. J., Jan., 1923, page 18, gives a diagram for the non-dissipative case of R, $K_{21}(m)$ and $K_{12}(m)/R$ in the transmitting band.

mean resistance, R, of the line with which the network is to be associated. If the network is closed at each end by a resistance of magnitude R, as in Fig. 2, we have not only a circuit arrangement which approximates more or less closely actual operating conditions,[5] but also a simple test circuit in which to determine the transmission loss of the network over the desired frequency range.

The transmission loss of a wave-filter network, defined with reference to Fig. 2, is the natural logarithm, with negative sign, of the ratio of the absolute value of the current transmitted from a source of resistance R to a receiving resistance R when the latter are connected through the network, to that transmitted when they are connected directly. Let E represent the electromotive force of the source, I the current transmitted to R through the network, and $E/2R$ that transmitted by direct connection. Then the *transmission loss* L, thus defined, is

$$L = -\log_e \left| \frac{I}{E/2R} \right|, \tag{1}$$

and

$$e^{-L} = \left| \frac{2RI}{E} \right|. \tag{2}$$

The unit in which L is expressed, the *attenuation unit*,[6] is the natural unit to use here and from the above relations it is seen that one attenuation unit of transmission loss corresponds to an absolute value of current ratio of $1 : e$. The method of determining the transmission loss under various possible conditions will be presented in the next part of this paper.

II. Theory of Chart Calculation Method

The principles given here are basic and apply to composite wave-filters having any terminations. However, in all practical cases, as previously stated, the terminations belong to the four types: 1, mid-shunt M-type; 2, mid-series M-type; 3, x-shunt "constant k;" and 4, x-series "constant k," all related to the "constant k" wave-filter.

[5] It should be clearly borne in mind that the unique selective properties of a wave-filter of freely transmitting currents in continuous frequency bands and of attenuating others are those for the wave-filter terminated in its characteristic impedance. It is practical to have approximately such a termination in the transmitting band only, as when connecting the wave-filter to a transmission line, in which case the general properties still persist. *Correct termination* rather than *number of sections* is what brings out these properties although the degree of selectivity is naturally increased by the addition of sections.

[6] A synonym sometimes used is the *Napier*. One attenuation unit is equivalent to 9.174 "800-cycle miles of standard cable," and to 8.686 TU. The TU (transmission unit) is that unit which designates a power ratio of 10^1, and the number of TU is ten times the common logarithm of the power ratio.

These four cases will be developed in detail and equivalence relations for certain sets of terminal combinations shown.

Fundamental Formula

The formula which is general and fundamental to what follows is the one giving the current received through a passive transducer in terms of the sending electromotive force, the terminal impedances,

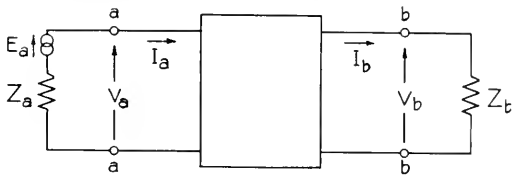


Fig. 3—General Linear Transducer

and the transfer constant [7] and image impedances of the transducer. Referred to Fig. 3 the received current is

$$I_b = \frac{2E_a \sqrt{W_a W_b}\, e^{-T}}{(W_a + Z_a)(W_b + Z_b)(1 - r_a r_b e^{-2T})}, \tag{3}$$

where

E_a = sending electromotive force,

Z_a, Z_b = sending and receiving impedances,

$T = D + iS$ = transfer constant of the transducer,

D, S = diminution constant and angular constant, defined as the real and imaginary parts of the transfer constant,

W_a, W_b = image impedances of the transducer at terminals a and b,

r_a, r_b = current reflection coefficients at terminals a and b,

$$r_a = \frac{W_a - Z_a}{W_a + Z_a},$$

and

$$r_b = \frac{W_b - Z_b}{W_b + Z_b}.$$

[7] The terms *transfer constant*, T and *image impedances*, W_a and W_b, as applied to a dissymmetrical passive transducer, are defined in the Appendix. These three parameters are to be distinguished from another set, the propagation constant, Γ, and characteristic impedances, K_a and K_b. In a symmetrical structure $T = \Gamma$ and $W_a = W_b = K_a = K_b$.

Another form obtained by suitable transformation is

$$I_b = \frac{E_a \sqrt{W_a W_b}}{(W_a W_b + Z_a Z_b)\sinh T + (W_a Z_b + W_b Z_a)\cosh T} \quad (4)$$

Formula (3), derived in the Appendix with several general transducer formulae and relations, is especially useful when applied to composite wave-filter networks, since, as we shall see, it contains the natural parameters for such structures. Upon comparing Fig. 2, which represents such a general network, with Fig. 3 we find that the two can be made to correspond exactly if the mid-part of the wave-filter, between terminals *a* and *b* in Fig. 2, is considered to be the transducer of Fig. 3, and if the wave-filter terminations combined with the resistances *R* are considered to be the terminal impedances Z_a and Z_b of Fig. 3. The relation between the electromotive force, E, applied in *R* and that, E_a, acting through Z_a depends upon the particular wave-filter termination at terminals *a*. Similarly, the relation between the currents, I and I_b, transmitted to *R* and Z_b, respectively, depends upon the termination at terminals *b*.

As already stated, the mid-part of the composite wave-filter consists in general of mid-series, mid-shunt, and mid-half sections, properly combined as to their impedance relations at the junction points. *The method of combination employed in a composite wave-filter consists in connecting two sections whose image impedances at their junction are equal.* (An analogy which might be given is the matching of dominoes in a line by the corresponding ends, numbers referring to image impedances.)

Let us assume for the moment that the mid-part, as thus made up, is terminated by impedances respectively equal to its image impedances. There is then an "image condition" for the impedances measured in the two directions not only at each of these terminal points but also at each junction point throughout the network; and in this case each section transmits under the "image condition" of its terminating impedances. As a result we obviously obtain the following properties for the mid-part.

1. *The transfer constant of the mid-part of a composite wave-filter, consisting of mid-series, mid-shunt, and mid-half sections, is the sum of the transfer constants of all the individual sections.*

2. *The image impedances of the mid-part of a composite wave-filter are the external image impedances of the two end sections.*

In addition we have the following important relations between the

transfer constant and image impedances of a single section, and the propagation constant and mid-point characteristic impedances of the corresponding ladder network.

3. *The transfer constant of a symmetrical mid-series or mid-shunt section is equal to the propagation constant of the corresponding ladder type; that of a dissymmetrical mid-half section (having mid-series and mid-shunt terminations) is equal to one-half the above propagation constant.*

4. *The image impedance of a mid-point section at a mid-series or mid-shunt point termination is equal to the mid-series, K_1, or mid-shunt, K_2, characteristic impedance, respectively, of the corresponding ladder network.*

Formula (3) is for the present purpose superior to the well known formula for transmitted current (derived for comparison in the Appendix) which contains the transducer recurrent parameters in the form of its propagation constant, Γ, and characteristic impedances, K_a and K_b. The reason for this is that in a dissymmetrical composite wave-filter where K_a differs from K_b, the usual case, no simple relations exist between these latter parameters of the transducer and the corresponding parameters of the individual sections comprising the network. In the special case of symmetrical networks, however, the latter formula becomes identical with (3) which follows from what has already been said.

Another method of obtaining the transmitted current, which may be termed the "section-by-section elimination method," consists in calculating by the aid of the Kirchhoff laws the current ratios and total impedances from section to section back through the entire network beginning at the receiving impedance. From the standpoint of time economy certain objections may be raised to the possible use here of this general long hand method of calculation. The method carries with it the determination of the phase as well as the amplitude of the transmitted current; but since the amplitude only is required in the transmission loss formula, this method does more than is necessary. Again, an alteration in the composite network structure requires a more or less complete recalculation when this method is employed, whereas by the application of (3) it will be found that this is not necessary. However, this method is useful where irregularities exist in the network, or where the particular method of design which had been followed in obtaining the composite structure cannot readily be found, but its impedance elements and R are known.

General Form of Transmission Loss Formula

Formulae (2) and (3) corresponding to Figs. 2 and 3 may be combined. If (3) is written in the general form

$$2RI\ E = F_t\ F_a\ F_b\ F_r, \tag{5}$$

we obtain with (2)

$$e^{\ L} = |2RI\ E| = e^{\ (L_t + L_a + L_b + L_r)}, \tag{6}$$

where the four factors comprising the current ratio $2RI\ E$ are

$F_t = e^{\ T}$ = the transfer factor between terminals a and b;

$F_a = \dfrac{2\sqrt{W_a R}\ E_a}{(W_a + Z_a)\ E}$ = the terminal factor at terminals a;

$F_b = \dfrac{2\sqrt{W_b R}\ I}{(W_b + Z_b)\ I_b}$ = the terminal factor at terminals b;

$F_r = \dfrac{1}{1 - r_a r_b e^{-2T}}$ = the interaction factor due to repeated reflections

at terminals a and b where the current reflection coefficients are

$$r_a = \frac{W_a - Z_a}{W_a + Z_a} \text{ and } r_b = \frac{W_b - Z_b}{W_b + Z_b};$$

and the transmission losses corresponding to the absolute values of these factors are called, respectively,

L_t = the transfer loss;

L_a, L_b = the terminal losses at terminals a and b;

and L_r = the interaction loss.

The total transmission loss is the sum of these four losses, thus,

$$L = L_t + L_a + L_b + L_r. \tag{7}$$

The relative importance of the three types of losses, transfer, terminal, and interaction, is usually in the order given. Hence, as a first approximation the transmission loss of a composite wave-filter is given by the transfer loss, L_t, but the error due to the omission of the other losses is often considerable. A second approximation is obtained by including the terminal losses, L_a and L_b, and for many purposes this is sufficiently accurate. The final step for accuracy is the further addition of the interaction loss, L_r, whose effect on the

total transmission loss is usually appreciable in the transmitting band of a wave-filter near the critical frequencies.

The three types of losses will now be considered separately and in detail.

1. *Transfer Losses*

The transfer loss, L_t, is by (6) equal to D, the diminution constant, which is the real part of the transfer constant, T, of the wave-filter mid-part taken between mid-points.

We have previously established the following:

(1) *T is the sum of the transfer constants of all the individual sections, i.e., $T=\Sigma Tj$; and* (2) *the transfer constant of a mid-series or mid-shunt section is equal to the propagation constant, $\Gamma=A+iB$ per full section, of the corresponding ladder type; that of a mid-half section is* Γ 2.

Hence, to get the transfer loss we need to know only the attenuation constant, A, of each full mid-section, the half or whole of which forms a part of the composite wave-filter structure. However, since the interaction factor which is to be discussed later requires a knowledge of the phase constant, B, as well, we shall consider both parts of the propagation constant at this point.

Propagation Constant of Ladder Type Network. The relation between the propagation constant $\Gamma=A+iB$, and the series and shunt impedances, z_1 and z_2, respectively, of the ladder type in Fig. 1 is known to be

$$\cosh \Gamma = 1+\frac{1}{2}\frac{z_1}{z_2}. \tag{8}$$

This applies as well to any recurrent structure if z_1 and z_2 correspond to the analytically equivalent ladder type.

Let us introduce two variables U and V by making the substitution

$$\frac{z_1}{4z_2} = U+iV. \tag{9}$$

The reason for this choice is that this ratio appears frequently in impedance formulae. Then in non-dissipative wave-filters, where $V=0$, the transmitting bands include all frequencies at which U satisfies the relation

$$-1 \leqq U \leqq 0. \tag{10}$$

By (8) and (9)

$$\cosh(A+iB)=\cosh A\cos B+i\sinh A\sin B=1+2U+i2V, \quad (11)$$

whence

$$\cosh A\cos B=1+2U,$$

and

$$\sinh A\sin B=2V. \quad (12)$$

The solution of this pair of simultaneous equations leads to separate relations for A and B,

$$\left(\frac{1+2U}{\cosh A}\right)^2+\left(\frac{2V}{\sinh A}\right)^2=1, \quad (13)$$

and

$$\left(\frac{1+2U}{\cos B}\right)^2-\left(\frac{2V}{\sin B}\right)^2=1. \quad (14)$$

As is well known from (13) equal attenuation constant loci are represented in the U, V plane by confocal ellipses with foci at $U=-1$, $V=0$ and $U=0$, $V=0$, thus having symmetry about the U-axis. *The locus for $A=0$, the limiting case, is a straight line between the foci and it corresponds to the transmitting band in a non-dissipative wave-filter.* Similarly from (14) equal phase constant loci are represented by confocal hyperbolas which have the same foci as above and are orthogonal to the equal attenuation constant ellipses. It will be assumed that the phase constant, B, lies between $-\pi$ and $+\pi$, which amounts to neglecting multiples of 2π. Then from (12) B has the same sign as V, so that loci in the upper half of the plane correspond to a positive phase constant while those in the lower half correspond to a negative one.

It is possible, however, to represent all this in just the upper half of the plane using coordinates U and $|V|$. Put

$$V=\epsilon|V|, \quad (15)$$

where $\epsilon=\pm 1$, the sign being that of V. The attenuation constant is independent of the sign of V, i.e., of ϵ. But for the phase constant we get from (12)

$$\sin \epsilon B=\frac{2|V|}{\sinh A}, \quad (16)$$

and

$$0\leqq \epsilon B\leqq +\pi.$$

Thus, as here considered, *the product ϵB, where $\epsilon=\pm 1$ has the sign of V, is always positive with a value less than or equal to π.*

Explicit formulae for A and B from (13) and (14) are

$$A = \sinh^{-1}\sqrt{2}\left[\sqrt{\sqrt{(U^2+U'^2+V'^2)^2+V'^2}+(U^2+U'^2+V'^2)}\right], \tag{17}$$

and

$$cB = \sin^{-1}\sqrt{2}\left[\sqrt{\sqrt{(U^2+U'^2+V'^2)^2+V'^2}-(U^2+U'^2+V'^2)}\right]. \tag{18}$$

The above formulae are general and applicable to any ladder type structure or its equivalent.

In the case of wave-filters certain approximate formulae are often useful. At frequencies in the attenuating bands away from the critical frequencies and the frequencies of maximum attenuation, and *wherever* V'^2 *is negligible compared with* $(U^2+U'^2)>0$,

$$A = \sinh^{-1} 2\sqrt{U^2+U'^2},$$

and (19)

$$cB = 0 \text{ or } \pi.$$

At the critical frequencies and the frequencies of maximum attenuation, *where* $(U^2+U'^2)$ *is negligible compared with* V'^2,

$$A = \cosh^{-1}(\sqrt{1+V'^2}+|V'|),$$

and (20)

$$cB = \cos^{-1}\pm(\sqrt{1+V'^2}-|V'|).$$

In the latter the positive sign applies to a critical frequency at which $U'=0$, and the negative sign to one at which $U'=-1$.

U' and V' for "Constant k" and M-type Wave-Filters. Since the wave-filter structures under consideration have "constant k" or derived M-type terminations, the U and V variables corresponding to these wave-filters will always be required. Hence, formulae for the variables are given here, limiting them to the four lowest wave-filter classes generally used.

Resistance in an inductance coil of inductance, L_1, is taken into account by expressing the total coil impedance as

$$(d+i)\ L_1 2\pi f,$$

where d, the "coil dissipation constant," is the ratio of coil resistance to coil reactance. The value of d is ordinarily between $d=.001$ and $d=.01$, and it does not vary rapidly with frequency. Similarly, dissipation in a condenser of capacity C_1 can be included by expressing the total condenser admittance as $(d'+i)\ C_1 2\pi f$, but since d' is usually negligible in practice it will here be omitted.

The formulae derived from (9) are based upon those given in this

Journal, Jan., 1923, pages 39 to 41, and contain the critical frequencies and frequencies of maximum attenuation. Subscripts $_k$ and $_m$ will be used to denote the "constant k" and M-type U and V variables. The "constant k" formulae for the four classes follow.

Low Pass.

$$U_k = -\left(\frac{f}{f_2}\right)^2,$$

and (21)

$$V_k = d\left(\frac{f}{f_2}\right)^2.$$

High Pass.

$$U_k = -\left(\frac{f_1}{f}\right)^2 / (1+d^2),$$

and (22)

$$V_k = -d\left(\frac{f_1}{f}\right)^2 (1+d^2).$$

Low-and-High Pass.

$$U_k = -\frac{(f_1-f_0)^2}{f_0 f_1}\frac{\left[\frac{f_0 f_1}{f^2} - (1+d^2)\left(2-\frac{f^2}{f_0 f_1}\right)\right]}{\left[\frac{f_0 f_1}{f^2} + (1+d^2)\frac{f^2}{f_0 f_1} - 2\right]^2},$$

and (23)

$$V_k = d\frac{(f_1-f_0)^2}{f_0 f_1}\frac{\left[\frac{f_0 f_1}{f^2} - (1+d^2)\frac{f^2}{f_0 f_1}\right]}{\left[\frac{f_0 f_1}{f^2} + (1+d^2)\frac{f^2}{f_0 f_1} - 2\right]^2}.$$

Band Pass.

$$U_k = -\frac{f_1 f_2}{(f_2-f_1)^2}\left[\frac{f_1 f_2}{(1+d^2)f^2} + \frac{f^2}{f_1 f_2} - 2\right],$$

and (24)

$$V_k = -d\frac{f_1 f_2}{(f_2-f_1)^2}\left[\frac{f_1 f_2}{(1+d^2)f^2} - \frac{f^2}{f_1 f_2}\right].$$

At the mid-frequency, $\sqrt{f_1 f_2}$, the point of confluency of two bands in the transmitting band of this wave-filter, we obtain approximately from (19), when d is small,

$$A = 2d\frac{\sqrt{f_1 f_2}}{f_2 - f_1},$$

and (25)

$$B = 0.$$

The derived M-type variables of any class are given directly in terms of the "constant k" variables of that class and the parameter m by the general relations

$$U_m = \frac{m^2[U_k + (1-m^2)(U_k^2 + V_k^2)]}{[1 + (1-m^2)U_k]^2 + (1-m^2)^2 V_k^2},$$

and (26)

$$V_m = \frac{m^2 V_k}{[1 + (1-m^2)U_k]^2 + (1-m^2)^2 V_k^2}.$$

This assumes that the M-type has the same grade of coils and condensers as its "constant k" prototype. The parameter m has a different formula determining its value for each class, the general relation being (neglecting dissipation)

$$m = \sqrt{1 + \left(\frac{1}{U_k}\right)_{f_\infty}}, \tag{27}$$

where f_∞ is a frequency of maximum attenuation of the M-type. The particular relations for the above four classes follow.

Low Pass

$$m = \sqrt{1 - \frac{f_a^2}{f_{2\infty}^2}}. \tag{28}$$

High Pass

$$m = \sqrt{1 - \frac{f_{1\infty}^2}{f_1^2}}. \tag{29}$$

Low-and-High Pass

$$m = \frac{\sqrt{\left(1 - \frac{f_0^2}{f_{1\infty}^2}\right)\left(1 - \frac{f_{1\infty}^2}{f_1^2}\right)}}{1 - \frac{f_0}{f_1}}. \tag{30}$$

Band Pass

$$m = \frac{\sqrt{\left(1 - \frac{f_1^2}{f_{2\infty}^2}\right)\left(1 - \frac{f_2^2}{f_{2\infty}^2}\right)}}{1 - \frac{f_1 f_2}{f_{2\infty}^2}}. \tag{31}$$

2. *Terminal Losses*

The general terminal losses L_a and L_b are determined by (6) from the absolute values of the terminal factors F_a and F_b, which factors we have assumed apply to the sending and receiving ends, respectively.

That either factor is dependent only upon its own type of termination and not upon its position at the sending or receiving end, can readily be shown. By the reciprocal theorem the product $F_t F_a F_b F_r$ is independent of the direction of current propagation, and from the forms of F_t and F_r the latter are also, whence the product $F_a F_b$ is independent of direction. Since in addition F_a and F_b are independent of each other they cannot depend upon position. This is equivalent to the statement that the ratios E_a / E and I / I_b which any particular termination would give at the sending and receiving ends, respectively, are equal. It will then be sufficient to consider the factor for a given termination at either end, say the receiving end.

The four terminations found practical give terminal losses which are reducible to two, namely, L_m and L_s now to be derived.

Terminal Losses, L_m, with Mid-M-type Terminations. These terminations, already mentioned, are

1. mid-shunt of a mid-series "constant k" equivalent M-type, $(K_{21}(m))$; and
2. mid-series of a mid-shunt "constant k" equivalent M-type, $(K_{12}(m))$.

The relations between the M-type characteristic impedances $K_{21}(m)$ and $K_{12}(m)$, the parameter m, and the variables U_k and V_k of the "constant k" prototype are, from formulae [8] in a previous paper

$$\frac{R}{K_{21}(m)} = \frac{K_{12}(m)}{R} = \frac{\pm \sqrt{1 + U_k \div i V_k}}{1 + (1 - m^2)(U_k \div i V_k)}. \tag{32}$$

Since $K_{12}(m) \cdot K_{21}(m) = R^2$, $K_{12}(m)$ and $K_{21}(m)$ are inverse networks of impedance product R^2. As either of these terminations is at a mid-point, it forms an end for the wave-filter mid-part and in the terminal factor F_b, arbitrarily chosen, $Z_b = R$ and $I / I_b = 1$, leaving

$$F_b = \frac{2 \sqrt{W_b R}}{W_b + R}. \tag{33}$$

In this factor the image impedance W_b is either $K_{21}(m)$ or $K_{12}(m)$, depending upon the type of termination. By (32) the factor is the same for both types provided they have the same parameter m, so

[8] The radicals which occur in this and succeeding formulae are proportional to physical impedances with positive resistance components. Hence, in each case the double sign is to be interpreted such as to make the real part of the radical positive.

that we may put for either of them *the single terminal loss* L_m defined by (6) as

$$e^{-L_m} = \left| \frac{2\sqrt{K_{21}(m)R}}{K_{21}(m)+R} = \frac{2\sqrt{K_{12}(m)R}}{K_{12}(m)+R} \right|$$

which upon the substitution of (32) gives

$$L_m = \log_e \left(\frac{1}{2} \left| 1 \pm \frac{\sqrt{1+U_k+iV_k}}{1+(1-m^2)(U_k+iV_k)} \cdot \left| \frac{1+(1-m^2)(U_k+iV_k)}{\sqrt{1+U_k+iV_k}} \right|^{\frac{1}{2}} \right| \right). \quad (34)$$

Terminal Losses, L_x, *with* x-"*constant* k" *Terminations.* The terminations are

3, x-shunt of the "constant k" wave-filter, (K_{x2}); and

4, x-series of the "constant k" wave-filter, (K_{x1}).

The x-shunt and x-series characteristic impedances, K_{x2} and K_{x1}, are related by the formulae

$$\frac{R}{K_{x2}} = \frac{K_{x1}}{R} = \frac{K_{1k}+(x-.5)z_{1k}}{R} = \pm\sqrt{1+U_k+iV_k} \pm (2x-1)\sqrt{U_k+iV_k}, \quad (35)$$

and

$$K_{x1}K_{x2} = K_{1k}K_{2k} = z_{1k}z_{2k} = R^2,$$

where K_{2k} and K_{1k} are the mid-shunt and mid-series values corresponding to $x = .5$. With either termination K_{x2} or K_{x1} it is assumed that the mid-part of the wave-filter begins at the mid-point, i.e., at the position corresponding to K_{2k} or K_{1k}, respectively, even when x is less than .5. In the latter case an impedance is theoretically added which is sufficient to "build-out" the wave-filter to the mid-point, and an equal impedance is similarly subtracted from the terminal impedance.

For termination 3, that is K_{x2}, the elements of factor F_b in (6) have the values

$$\begin{aligned} II_b &= K_{2k}, \\ Z_b &= z_{2k}\, R\ (z_{2k}+(x-.5)R), \\ I\ \ I_b &= z_{2k}\ (z_{2k}+(x-.5)R). \end{aligned} \quad (36)$$

and

For termination 4, K_{x1}, they are

$$\begin{aligned} II_b &= K_{1k}, \\ Z_b &= R+(x-.5)z_{1k}, \\ I\ \ I_b &= 1. \end{aligned} \quad (37)$$

and

The substitution of (36) or (37) in F_b gives an identical result, as shown by relations (35), provided x is the same in both. *A single terminal loss L_x may then apply to either*, which is defined from (6) as

$$e^{-L_x} = \frac{2\sqrt{(R^2 K_{2k})R}}{R^2 K_{x2} + R} = \frac{2\sqrt{K_{1k}R}}{K_{x1} + R},$$

giving by (35)

$$L_x = \log_e\left(\frac{1}{2}\left|1 \pm \sqrt{1 + U_k + iV_k} \pm (2x-1)\sqrt{U_k + iV_k}\right| \cdot \left|1 + U_k + iV_k\right|^{\frac{1}{4}}\right). \quad (38)$$

A comparison of (34) and (38) shows that when $m = 1$ and $x = .5$, $L_m = L_x$ as should be the case.

3. *Interaction Losses*

The interaction loss defined in (6) is expressible in its general form as

$$L_r = \log_e \left|1 - r_a r_b e^{-2T}\right|. \quad (39)$$

It depends not only upon the transfer constant T, including both diminution and angular constants, but also upon the complex reflection coefficients, r_a and r_b, at the two ends. That is, it is a function both of the internal structure and of the terminations of the wave-filter. For this reason its determination offers the most complexity of all the three types of losses and, in fact, requires a knowledge of the transfer loss. On the other hand, it is usually the least important part of the total transmission loss and may usually be omitted except at frequencies within a transmitting band and near a critical frequency.

The transfer constant $T = D + iS$ is given by the relations and formulae developed when considering the transfer loss.

The multiplication of the reflection coefficients and the square of the transfer factor is simplified to a problem in addition by expressing each of these coefficients in the exponential form,

$$r_a = e^{-G_a - iH_a},$$

and

$$r_b = e^{-G_b - iH_b}.$$

Then, putting $r_a r_b e^{-2T} = e^{-P - iQ}$,

$$L_r = \tfrac{1}{2}\log_e (1 + e^{-2P} - 2e^{-P}\cos Q), \quad (40)$$

where $P = G_a + G_b + 2D,$

and $Q = H_a + H_b + 2S.$

The subscripts, as before, merely refer to the terminations. The G and H expressions which correspond to the reflection coefficient with each of the four particular types of terminations, 1, 2, 3, and 4, follow.

Reflection Coefficients, r_{m2} and r_{m1}, with Mid-M-type Terminations. For termination 1 arbitrarily assumed at b we have $W_b = K_{21}(m)$ and $Z_r = R$. Introducing for this case the subscript $_{m2}$, signifying M-type and mid-shunt, it follows by (6) and (32) that

$$r_{m2} = \frac{K_{21}(m) - R}{K_{21}(m) + R},$$

and its equivalent

$$e^{-G_{m2} - iH_{m2}} = r_{m2} = \frac{1 + (1 - m^2)(U_k + iV_k) - (\pm\sqrt{1 + U_k + iV_k})}{1 + (1 - m^2)(U_k + iV_k) \pm \sqrt{1 + U_k + iV_k}}. \quad (41)$$

With termination 2, $W_b = K_{12}(m)$ and $Z_b = R$, so that by (32) the corresponding coefficient r_{m1} becomes

$$r_{m1} = -r_{m2},$$

or

$$e^{-G_{m1} - iH_{m1}} = -e^{-G_{m2} - iH_{m2}}. \quad (42)$$

Since $-1 = e^{i\pi}$,

$$G_{m1} = G_{m2},$$

and

$$H_{m1} = H_{m2} + \pi. \quad (43)$$

Reflection Coefficients, r_{x2} and r_{x1}, with x-"constant k" Terminations. In the case of the x-shunt termination 3, K_{x2}, relations (36) give

$$r_{x2} = \frac{K_{2k} - z_{2k}R \ (z_{2k} + (x - .5)R)}{K_{2k} + z_{2k}R \ (z_{2k} + (x - .5)R)}.$$

Introducing (35) this is

$$e^{-G_{x2} - iH_{x2}} = r_{x2} = \frac{1 \pm (2x - 1)\sqrt{U_k + iV_k} - (\pm\sqrt{1 + U_k + iV_k})}{1 \pm (2x - 1)\sqrt{U_k + iV_k} \pm \sqrt{1 + U_k + iV_k}}. \quad (44)$$

The x-series termination 4, K_{x1}, has a coefficient r_{x1} determined by (37) which is related to r_{x2} through (35) as

$$r_{x1} = -r_{x2}. \quad (45)$$

It follows from the corresponding exponential expressions that

$$G_{x1} = G_{x2},$$

and

$$H_{x1} = H_{x2} + \pi. \quad (46)$$

Hence, the two members of each pair of reflection coefficients, r_{m2}, r_{m1}, and r_{x2}, r_{x1}, differ only in sign so that their G's are the same but their H's differ by π.

I. Wave-Filter Structures Having Equivalent Transmission Losses

There are six groups of possible wave-filter networks involving the four terminations above, each group of which is made up of pairs having equivalent current ratios $2RI\ E$ and hence equivalent transmission losses. By (5) this means that the members of such a pair have products for their four factors, $F_1F_aF_bF_r$, which are equal. It may readily be shown from preceding relations that these groups, represented symbolically by brackets enclosing the transfer constants of their mid-parts and the terminations, are the following:

$$\begin{aligned}
&(a)\ [T, K_{21}(m), K_{21}(m')] = [T, K_{12}(m), K_{12}(m')],\\
&(b)\ [T, K_{21}(m), K_{12}(m')] = [T, K_{12}(m), K_{21}(m')],\\
&(c)\ [T, K_{21}(m), K_{x2}] \quad = [T, K_{12}(m), K_{x1}],\\
&(d)\ [T, K_{21}(m), K_{x1}] \quad = [T, K_{12}(m), K_{x2}],\\
&(e)\ [T, K_{x2}, K_{x}'_2] \quad = [T, K_{x1}, K_{x}'_1],\\
&(f)\ [T, K_{x2}, K_{x}'_1] \quad = [T, K_{x1}, K_{x}'_2].
\end{aligned} \tag{17}$$

This symbolic representation in (*c*), for example, means that a composite wave-filter whose mid-part has a transfer constant, T, and whose terminations are those designated by $K_{21}(m)$ and K_{x2}, will give the same current ratio $2RI\ E$ as another wave-filter whose mid-part has the same transfer constant, T, but whose terminations are those designated by $K_{12}(m)$ and K_{x1} where m and x are respectively the same in both networks.

III. Charts for Determining Transmission Losses

The accompanying charts apply to the three groups of transmission losses, transfer, terminal, and interaction, and are derived from the general formulae already given. The curves represent constant parameter loci for A, cB, L_m, L_x, G_{m2}, cH_{m2}, G_{x2}, cH_{x2}, and L_r as functions of several variables and include the most practical range; where further extension is required the original formulae may be consulted. The U and V variables for the ladder type of recurrent network (or its equivalent) which form the basis of this chart calculation method are to be found as a function of frequency, in the general case from formula (9),

$$z_1\ kz_2 = U + iV,$$

and in the lower class "constant k" and M-type wave-filters from formulae (21) to (31). Owing to the large number of intermediate equations which it was necessary to obtain before direct computations could suitably be made for the charts, these equations will not be given here, but only a brief designation of the resulting charts together with the approximations involved, if any.

The units employed throughout are the *attenuation unit* and the *radian*. The former unit applies to A, D, L_m, L_x, G_{m2}, G_{x2}, P, L_r and L, and the latter unit to B, S, H_{m2}, H_{x2} and Q.

Transfer Loss

This is determined through the propagation constant, $\Gamma = A + iB$.

Charts 1, 2, *and* 3.—A *and* cB *in and about transmitting band*; $c = \pm 1$ has the sign of U.

Chart 4.—A *in attenuating band*; U^2 negligible compared with $(U + U^2) > 0$.

Chart 5.—A *at maximum attenuation*; $(U + U^2)$ negligible compared with U^2.

Terminal Losses, L_m and L_x

Chart 6.—L_m *in transmitting band*; U_k neglected.

Chart 7.—L_m *at critical frequency*; $U_k = -1$.

Chart 8.—L_m *in attenuating band*; U_k neglected.

Chart 9.—L_m *at maximum attenuation of M-type*; $U_k = -\frac{1}{1 - m^2}$.

Chart 10.—L_x *in transmitting band*; U_k neglected.

Chart 11.—L_x *at critical frequency*; $U_k = -1$.

Chart 12.—L_x *in attenuating band*; U_k neglected.

Reflection Coefficients

Note that

$$G_{m1} = G_{m2},$$

and $$H_{m1} = H_{m2} + \pi;$$

also that

$$G_{x1} = G_{x2},$$

and $$H_{x1} = H_{x2} + \pi.$$

Chart 13. *G_{m2} and H_{m2} in transmitting band*;
V_k neglected.

Chart 14. *G_{m2} and cH_{m2} at critical frequency*;
$V_k = -1$ and $c = \pm 1$ has the sign of V_k.

Chart 15. *G_{m2} and cH_{m2} in attenuating band*;
V_k neglected.

Chart 16. *G_{x2} and cH_{x2} in transmitting band*;
V_k neglected and $c = \pm 1$ has the sign of V_k.

Chart 17. *G_{x2} and cH_{x2} at critical frequency*;
$V_k = -1$.

Chart 18. *G_{x2} and cH_{x2} in attenuating band*;
V_k neglected.

Interaction Loss, L_r

Note that $T = D + iS$ = transfer constant of mid-part of wave-filter;

$$P = G_a + G_b + 2D,$$

and $$Q = H_a + H_b + 2S,$$

where a and b refer to the terminations.

Chart 19.—*L_r as a function of P and Q.*

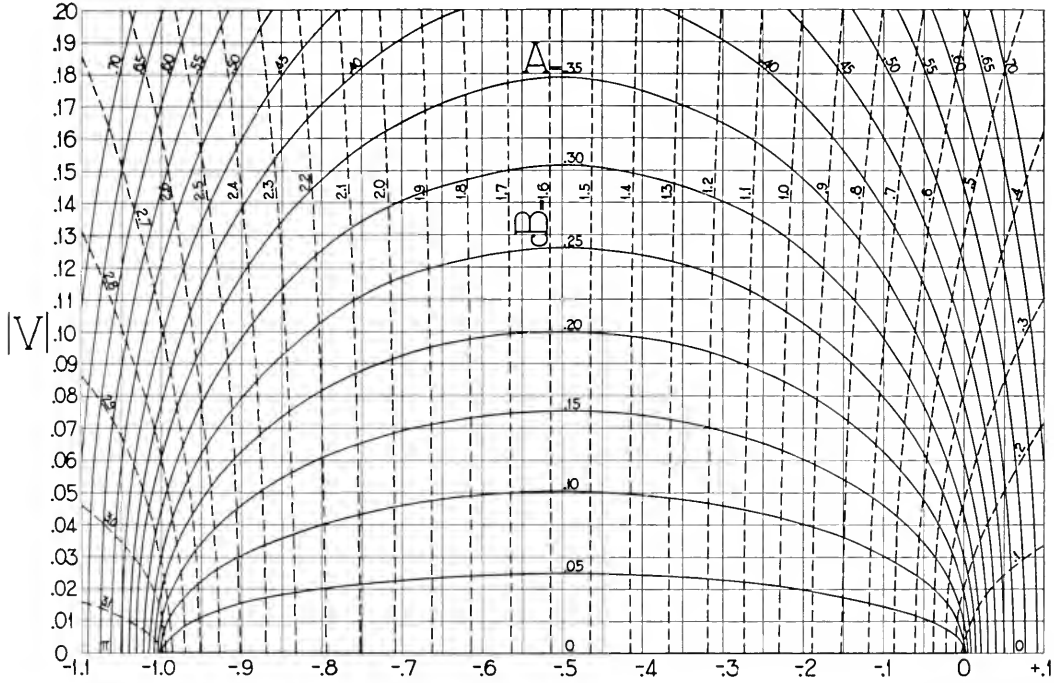


Chart 1.—A and cB in and about transmitting band;

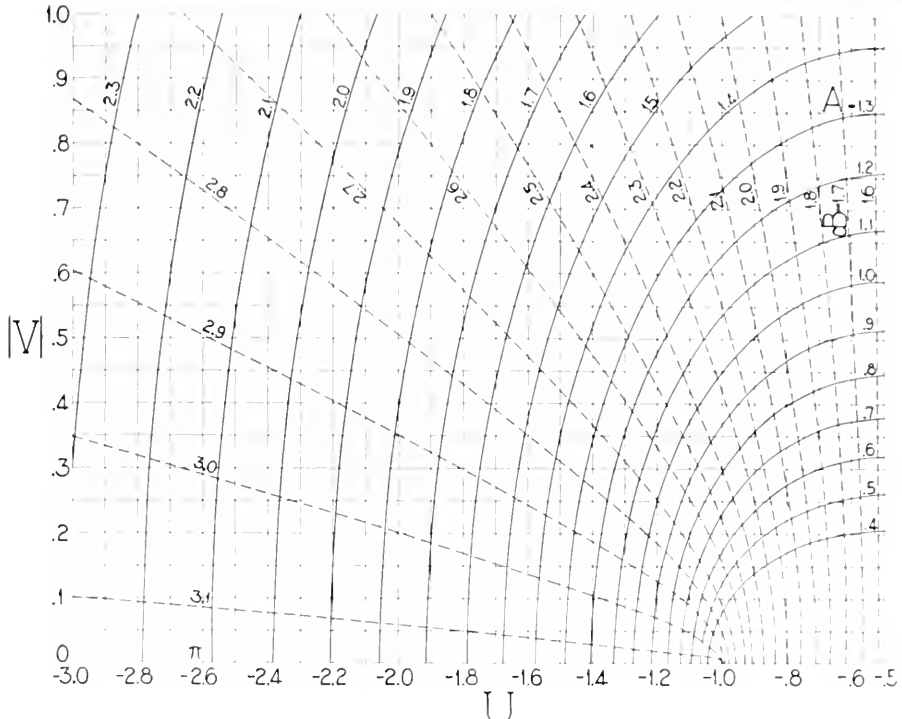


Chart 2. —*A and cB in and about transmitting band,* $c = \pm 1$ has the sign of *U*.

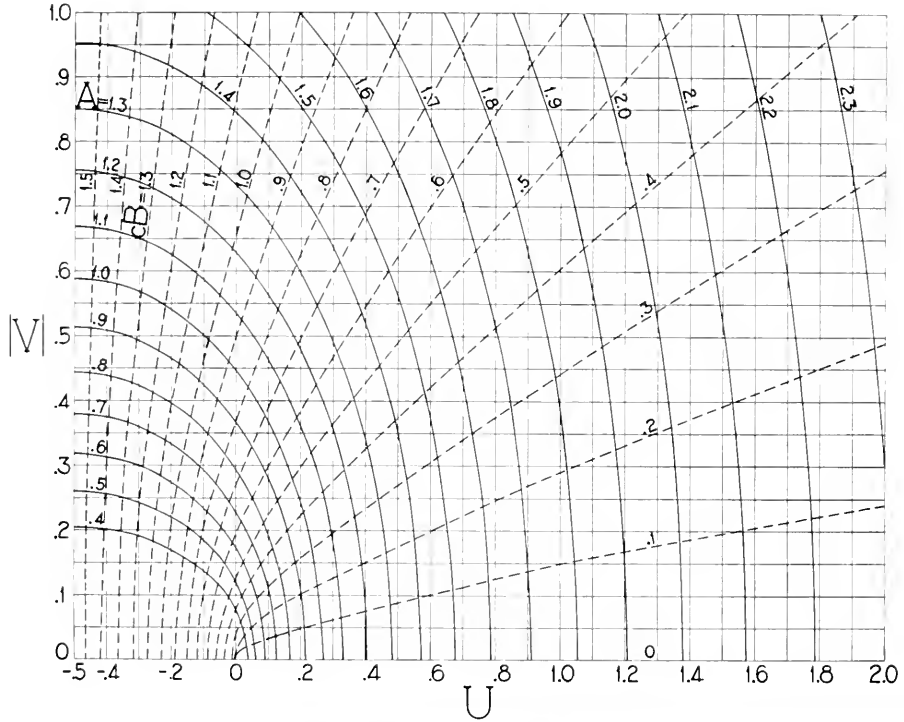


Chart 3.—A and cB in and about transmitting band;
$c = \pm 1$ has the sign of V.

50 55 60 65 70 75 80 85 90 95 100
-71 -76 -81 -86 -91 -96 -101

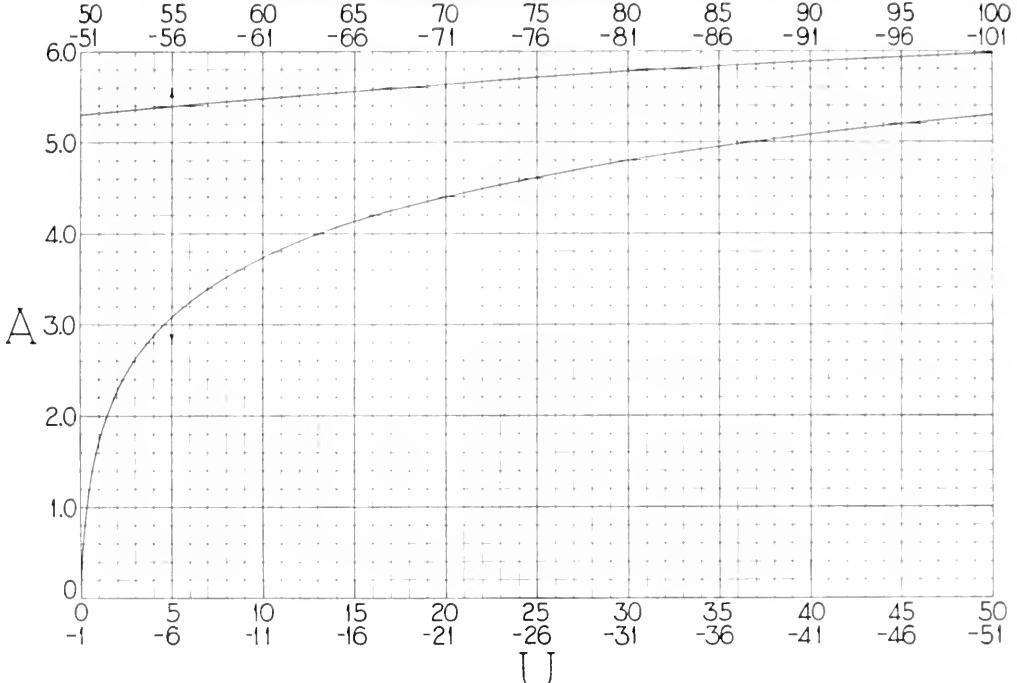


Chart I — A in attenuating band,
U^2 negligible compared with $(U + U^2 + \ldots)$.

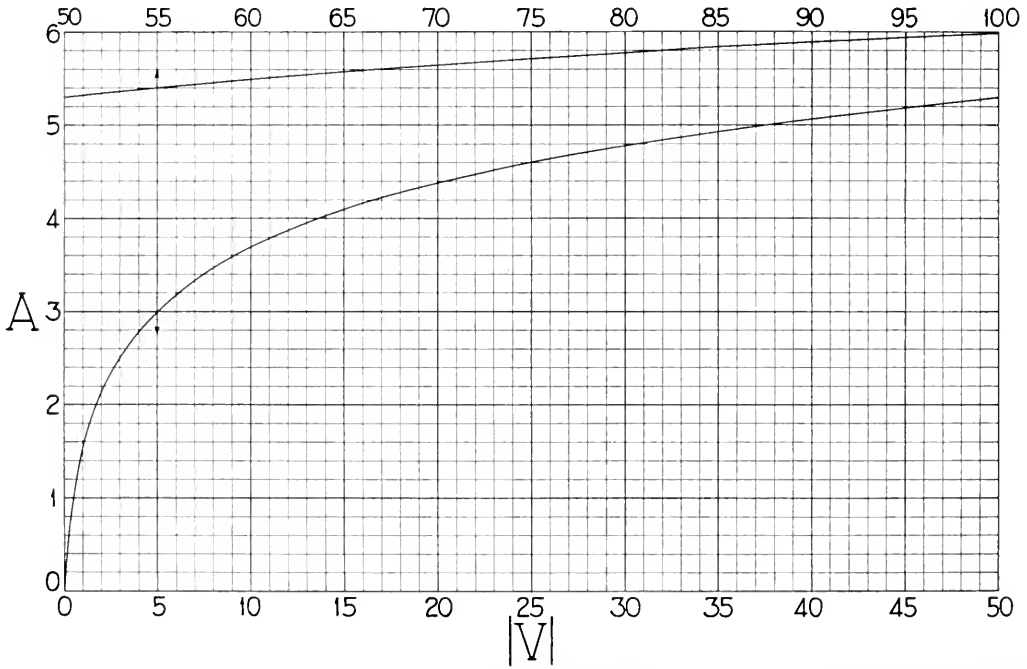


Chart 5.—A at maximum attenuation;
$(U+U^2)$ negligible compared with 1^2.

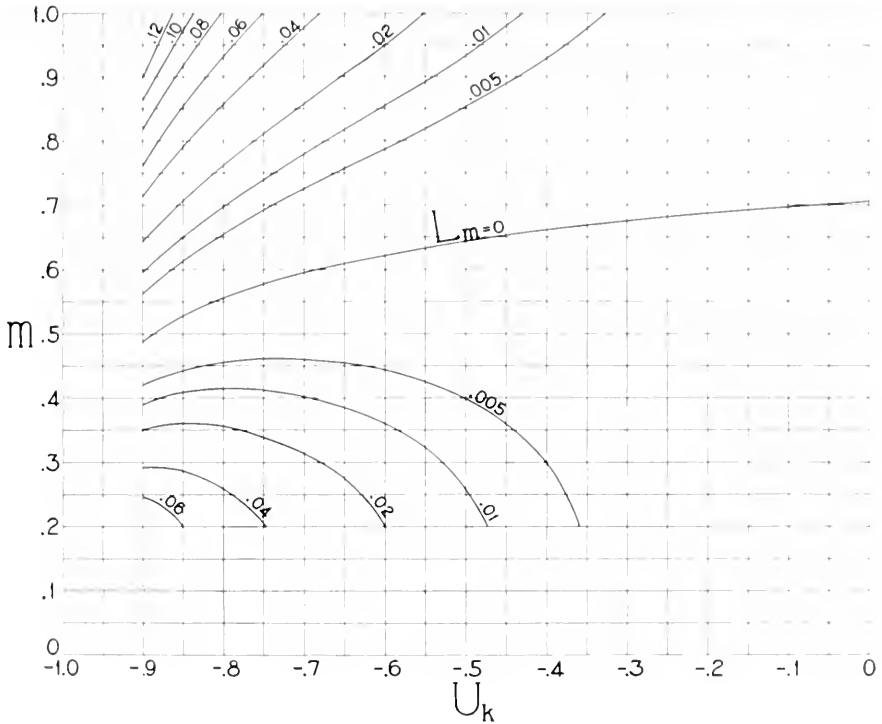


Chart 6. L_m in transmitting band; V_k neglected.

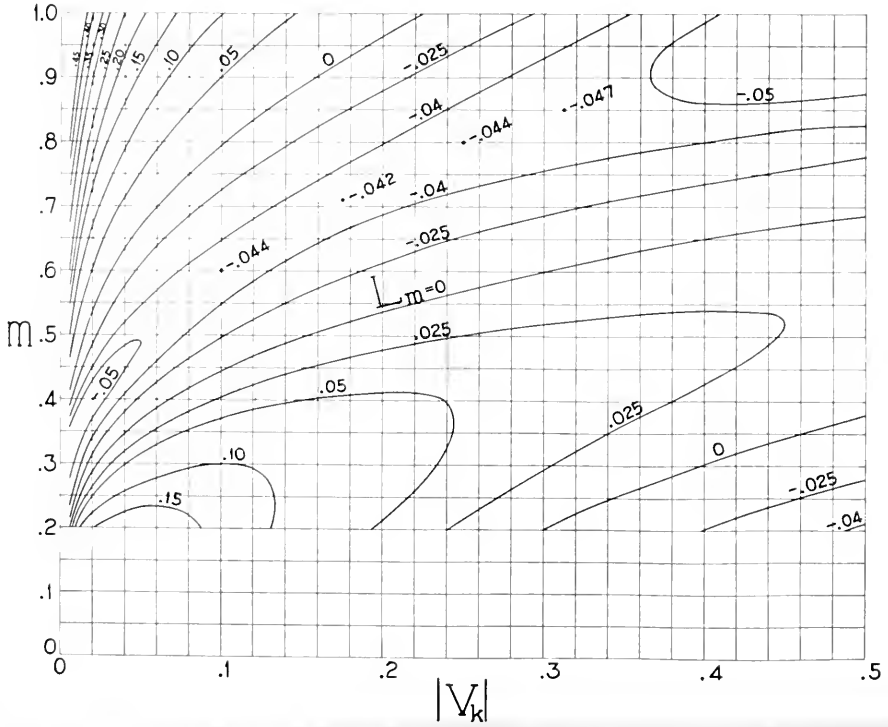


Chart 7.—L_m at critical frequency.

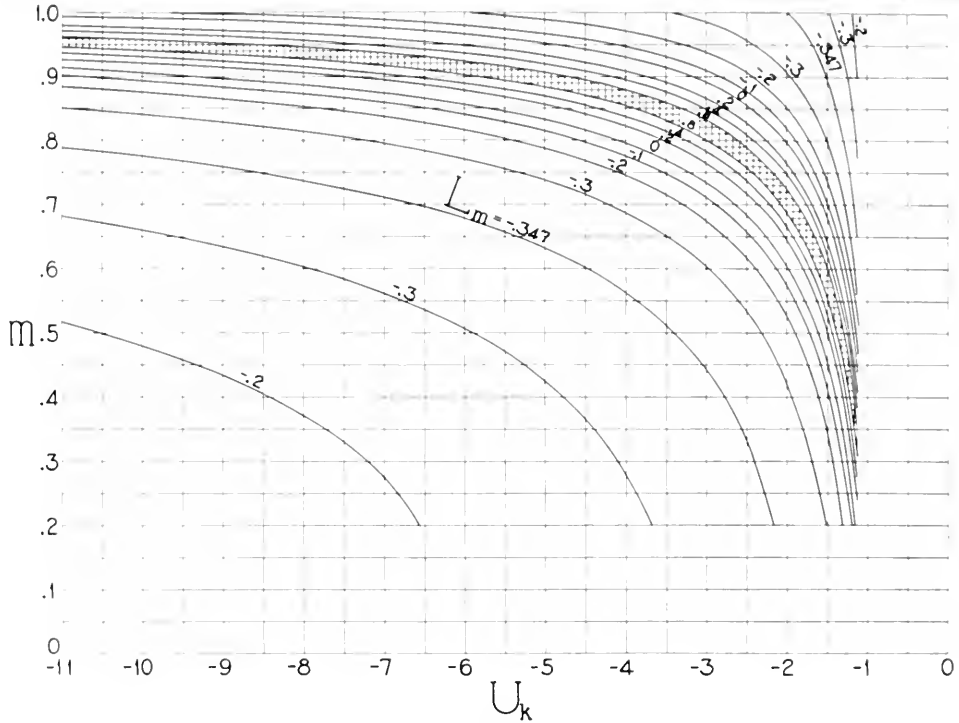


Chart 8 L_m *in attenuating band,* V_k neglected.

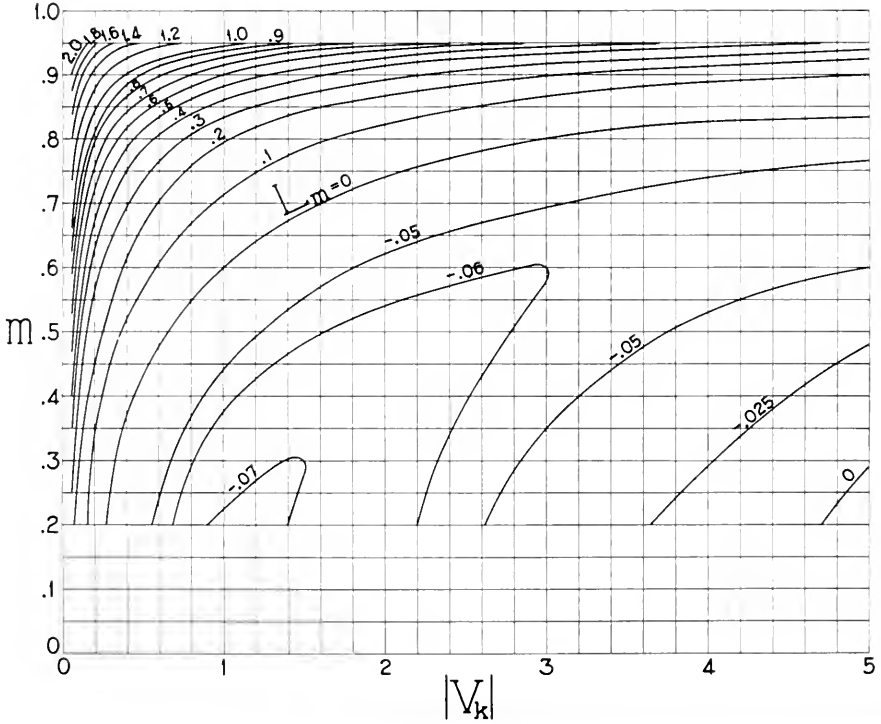


Chart 9.—L_m at maximum attenuation of M-type;

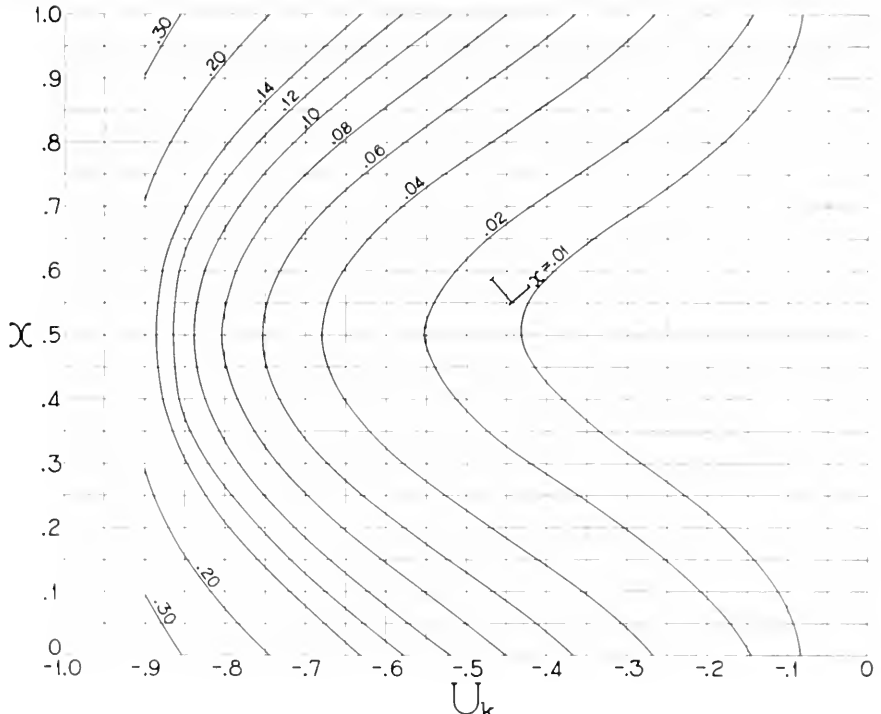


Chart 10. L_x in transmitting band, V_k neglected.

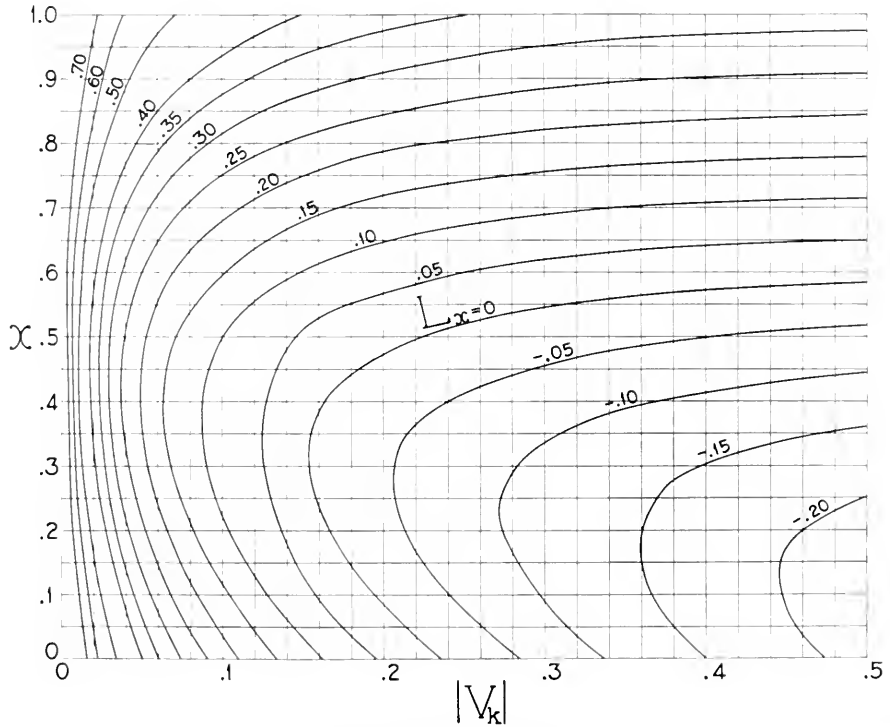


Chart 11.– L_x at critical frequency; $U_k = -1$.

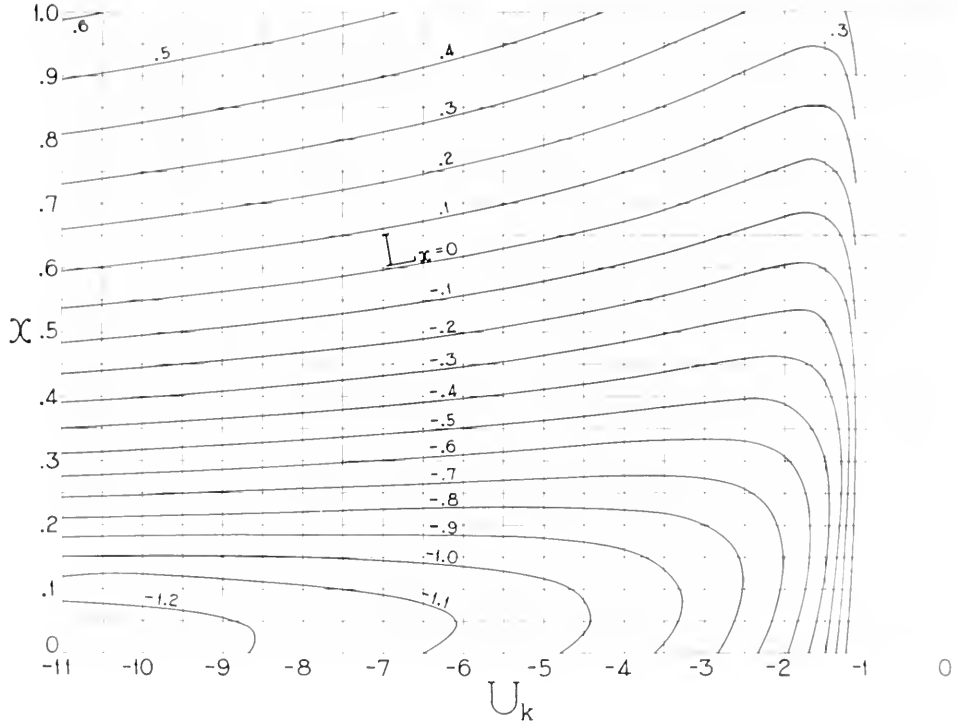


Chart 12. L_x in attenuating band.

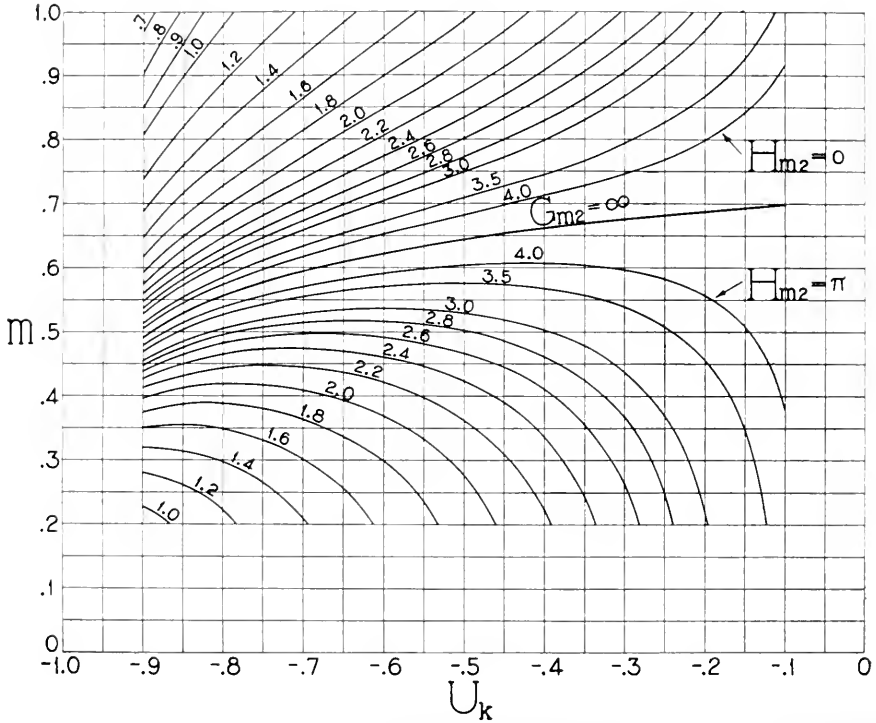


Chart 13.—G_{m2} *and* H_{m2} *in transmitting band;* V_k neglected.

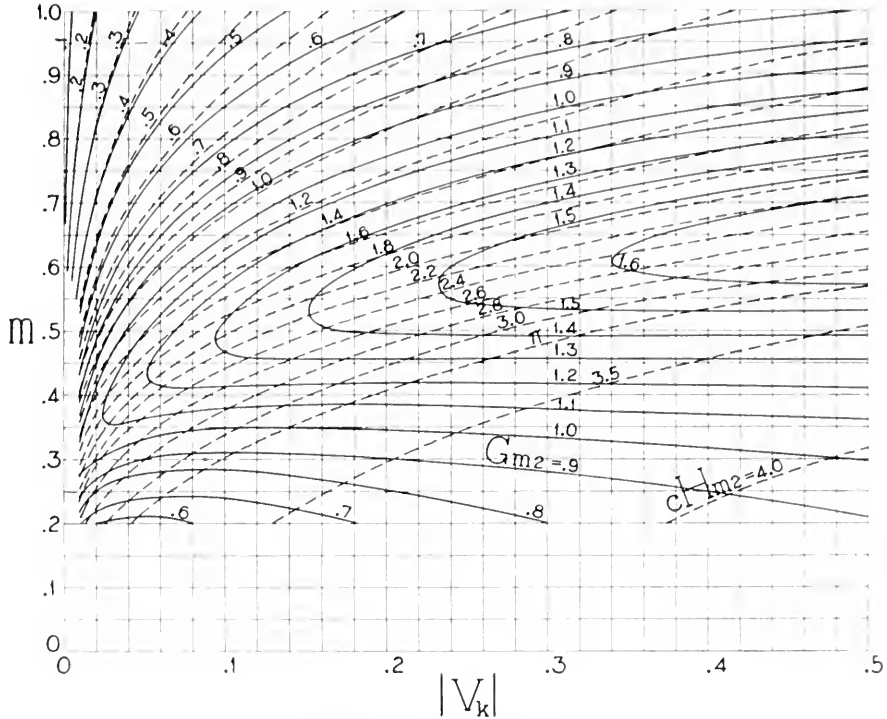


Chart 14 — G_{m2} and cH_{m2} at critical frequency; $V_k = -1$ and $c = +1$ has the signo f V_k

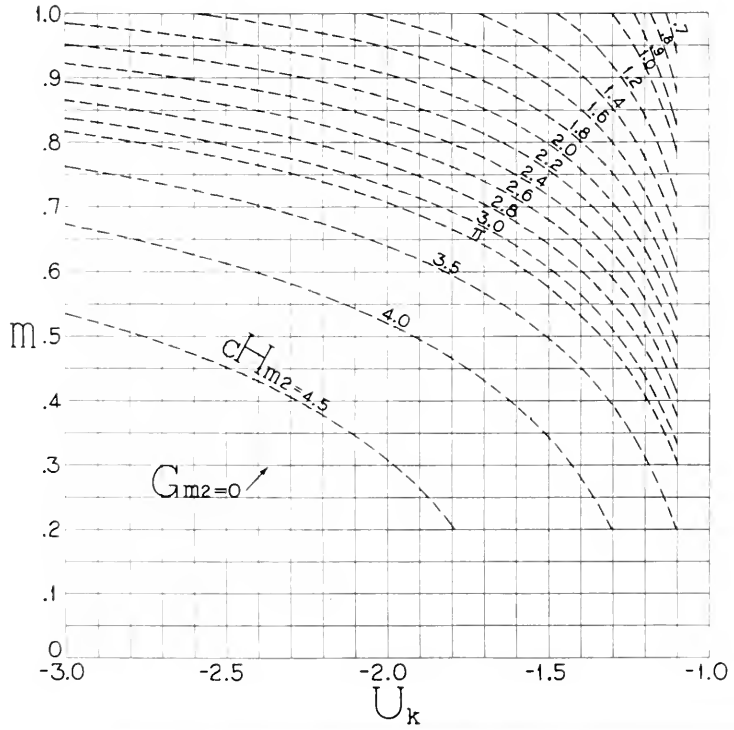


Chart 15.—G_{m2} and cH_{m2} in attenuating band;

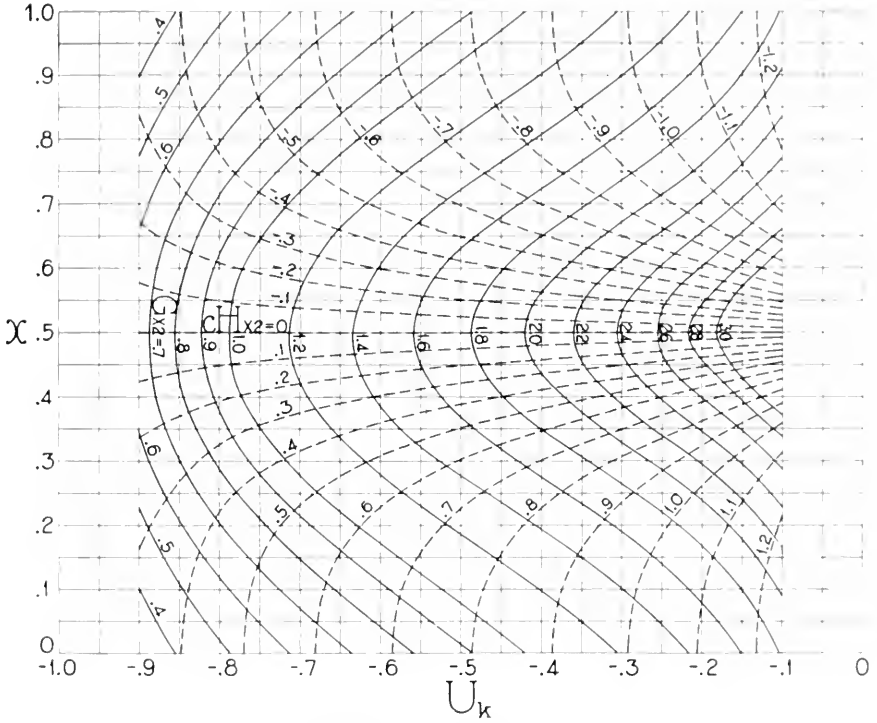


Chart 16. G_{x2} and dH_{x2} in transmitting band;

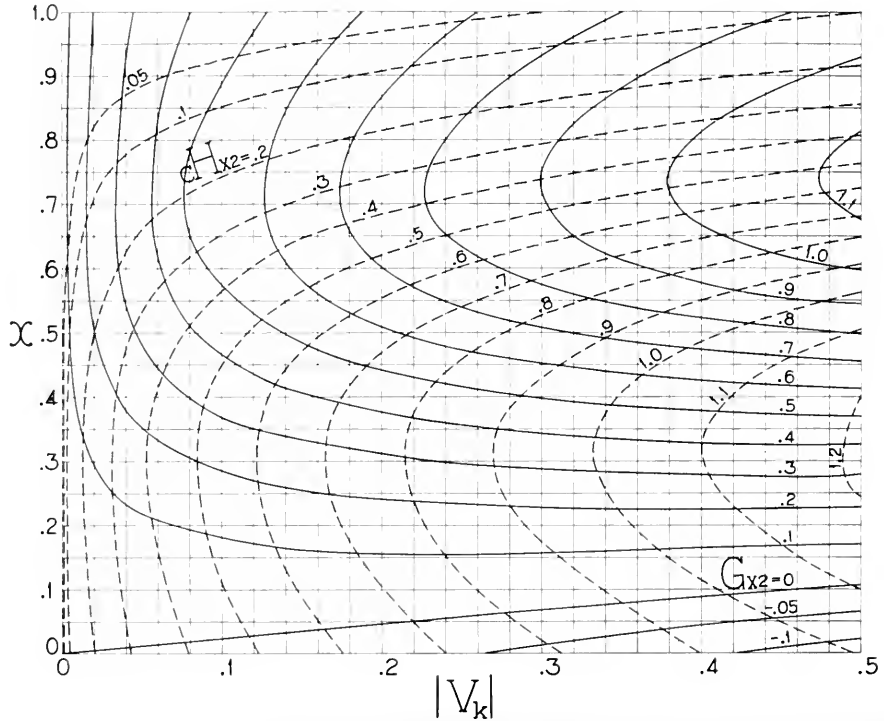


Chart 17.—G_{x2} and cH_{x2} at critical frequency; $U_k = -1$.

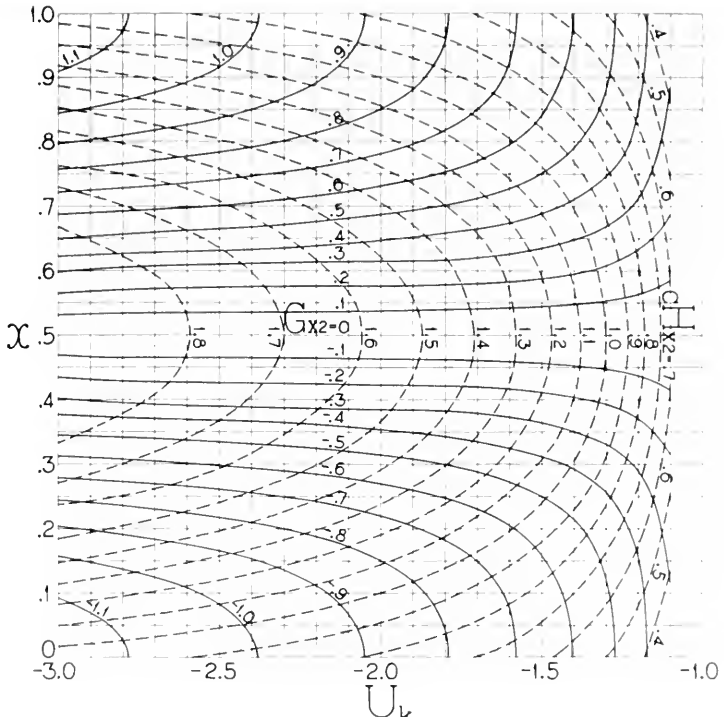


Chart 18 — G_{x2} and dH_{x2} in attenuating band, U_k neglected.

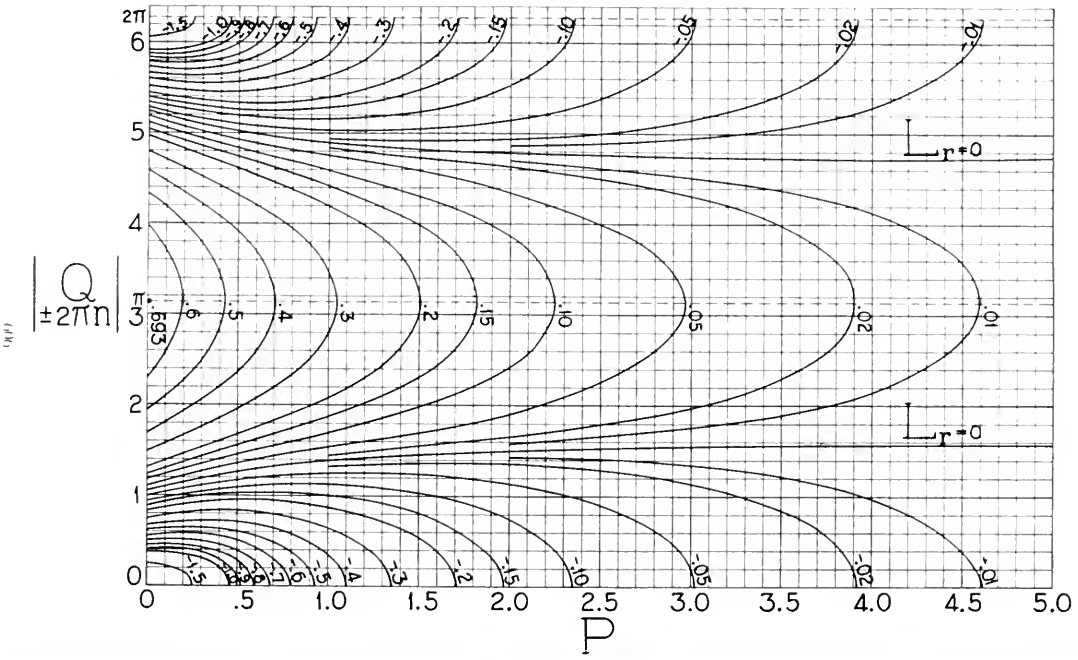


Chart 19. L_r as a function of P and Q.

IV. Illustrative Application of the Method

In the following illustration a small number of band pass wave-filter sections having different characteristics is chosen purposely so as to allow an appreciable interaction factor and the use of all the charts.

The mid-part of the composite wave-filter is made up of one mid-series section of type VI_1 and one mid-half section of M-type IV_1, the designations being those of a previous paper. The termination at one end is made K_{c1} by adding $(x-.5)z_{1k}$ in series with type VI_1 and at the other is $K_{c2}(m)$, as is diagrammatically represented at the top of Fig. 4. The values of all the parameters were chosen as follows:

$R = 600$ ohms, $\qquad x = .80,$

$f_1 = 4{,}000 \sim,$ $\qquad M$-type, $f_{2\infty} = 8{,}000 \sim,$

$f_2 = 7{,}000 \sim,$

and $d = .01$ (assumed constant for computation purposes).

With these values the magnitudes and locations of inductances and capacities are as shown in the center of Fig. 4, where the series impedance parts have been merged together.

The variables U_k and V_k for the "constant k" band pass wave-filter as well as U_m, V_m, and m of the M-type are given by formulae (21), (26), and (31). In the 3-element type VI_1

$$U_1 + iV_1 = \frac{1-(f/f_1)^2}{(f_2/f_1)^2-1} + id\frac{(f/f_1)^2}{(f_2/f_1)^2-1} \qquad (18)$$

These variables have been computed in the present case for frequencies on both sides of the transmitting band and are tabulated below. The other tables including that of transmission losses are based upon this table and the charts.

The next to the last, and the last columns give the total transmission losses as obtained by this chart method and by direct network computation, respectively. Comparison shows that there is a very satisfactory agreement between them, the differences at all frequencies being negligible in practice. The greatest differences of approximately .05 attenuation units at frequencies 3750 and 7500 cycles per second, just outside the transmitting band, can readily be explained as due to the omission of dissipation in the two terminal loss factors and the reflection coefficients. The transmission loss is shown graphically at the bottom of Fig. 4.

It is believed that the use of this chart method will result in considerable time economy with calculations of this kind.

TABLE I
U and V Variables

f Cycles sec.	"Constant k"		Type VI$_1$		M-type, $m=.7454$	
	U_k	V_k	U_1	V_1	U_m	V_m
1000	−81.0	−.870	.455	.0003	1.29	−.0004
1500	−32.7	−.385	.417	.0007	1.34	−.0012
2000	−16.0	−.213	.364	.0012	1.45	−.0032
2500	−8.41	−.132	.296	.0019	1.71	−.0098
3000	−4.46	−.0868	.212	.0027	2.53	−.050
3250	−3.20	−.0707	.165	.0032	4.21	−.220
3500	−2.25	−.0575	.114	.0037	.36	−48.9
3750	−1.53	−.0463	.0586	.0042	−2.67	−.253
4000	−1.00	−.0367	0	.0049	−.997	−.0659
4250	−.607	−.0282	−.0625	.0055	−.461	−.0293
4500	−.329	−.0205	−.129	.0061	−.214	−.0156
5292	.00031	0	−.364	.0085	.00017	0
6500	−.534	.0263	−.796	.0128	−.389	.0252
6750	−.752	.0315	−.897	.0138	−.627	.0395
7000	−1.00	.0367	−1.00	.0148	−.998	.0659
7500	−1.58	.0470	−1.22	.0170	−2.90	.290
8000	−2.25	.0575	−1.46	.0194	.85	48.9
8500	−3.01	.0682	−1.70	.0219	4.92	.329
9000	−3.85	.0792	−1.97	.0246	3.00	.0866
10000	−5.76	.102	−2.55	.0303	2.05	.0234
11000	−7.94	.127	−3.18	.0367	1.75	.0110
12000	−10.4	.154	−3.88	.0436	1.60	.0065

TABLE II
Transfer Constants

f Cycles sec.	Mid-series Type VI$_1$		Mid-half M-type IV_1		Mid-part of Wave-filter	
	$T_1=A_1+iB_1$		$T_m=\frac{1}{2}(A_m+iB_m)$		$T=T_1+T_m=D+iS$	
1000	1.20	—	.97	—	2.23	—
1500	1.21	—	.99	—	2.20	—
2000	1.14	—	1.02	—	2.16	—
2500	1.04	—	1.08	—	2.12	—
3000	.89	—	1.23	—	2.12	—
3250	.79	—	1.46	—	2.25	—
3500	.66	—	2.64	—	3.30	—
3750	.480	+i .02	1.077	−i1.51	1.557	−i1.49
4000	.100	+i .10	.181	−i1.39	.281	−i1.29
4250	.025	+i .51	.029	−i .75	.054	−i .24
4500	.019	+i .73	.019	−i .48	.038	+i .25
5292	.018	+i1.30	.013	+i 0	.031	+i1.30
6500	.032	+i2.20	.026	+i .67	.058	+i2.87
6750	.043	+i2.48	.040	+i .92	.083	+i3.40
7000	.173	+i2.97	.181	+i1.39	.354	+i4.36
7500	.910	+i3.10	1.125	+i1.51	2.035	+i4.61
8000	1.27	—	2.64	—	3.91	—
8500	1.52	—	1.53	—	3.05	—
9000	1.74	—	1.31	—	3.05	—
10000	2.09	—	1.15	—	3.24	—
11000	2.36	—	1.09	—	3.45	—
12000	2.59	—	1.06	—	3.65	—

TABLE III

Reflection Coefficients and Interaction Factor

f Cycles sec.	$x = .80$		$m = .7454$		P	Q
	G_{x1}	H_{x1}	G_{m2}	H_{m2}		
3750	.58	2.18	0	−2.31	3.69	−3.11
4000	.30	3.06	.49	−.50	1.35	−.02
4250	1.04	3.75	2.58	0	3.73	3.27
4500	1.57	4.01	3.85	0	5.50	4.54
5292	∞	−	∞	−	∞	
6500	1.17	2.45	2.90	0	4.19	8.19
6750	.78	2.67	1.92	0	2.87	9.47
7000	.30	3.22	.49	.50	1.50	12.44
7500	.60	4.14	0	2.39	4.67	15.75

TABLE IV

Transmission Losses

f Cycles sec.	Transfer	Terminal		Interaction	Total = L	
	L_t	L_x	L_m	L_r	ΣL_j	Network Computation
1000	2.23	.88	.02	—	3.13	3.13
1500	2.20	.65	−.18	—	2.67	2.68
2000	2.16	.48	−.30	—	2.34	2.35
2500	2.12	.33	−.35	—	2.10	2.11
3000	2.12	.19	−.25	—	2.06	2.08
3250	2.25	.12	−.01	—	2.36	2.37
3500	3.30	.06	1.21	—	4.57	4.59
3750	1.557	.042	−.190	.025	1.434	1.487
4000	.281	.443	.082	−.300	.506	.508
4250	.054	.067	.004	.024	.149	.154
4500	.038	.023	.001	.001	.063	.068
5292	.031	.000	.000	.000	.031	.036
6500	.058	.052	.003	.005	.118	.127
6750	.083	.118	.011	.055	.267	.276
7000	.354	.443	.082	−.250	.629	.632
7500	2.035	.038	−.150	.009	1.932	1.987
8000	3.91	.06	1.21	—	5.18	5.19
8500	3.05	.11	.06	—	3.22	3.24
9000	3.05	.16	−.18	—	3.03	3.05
10000	3.24	.24	−.31	—	3.17	3.18
11000	3.45	.31	−.34	—	3.42	3.43
12000	3.65	.38	−.33	—	3.70	3.70

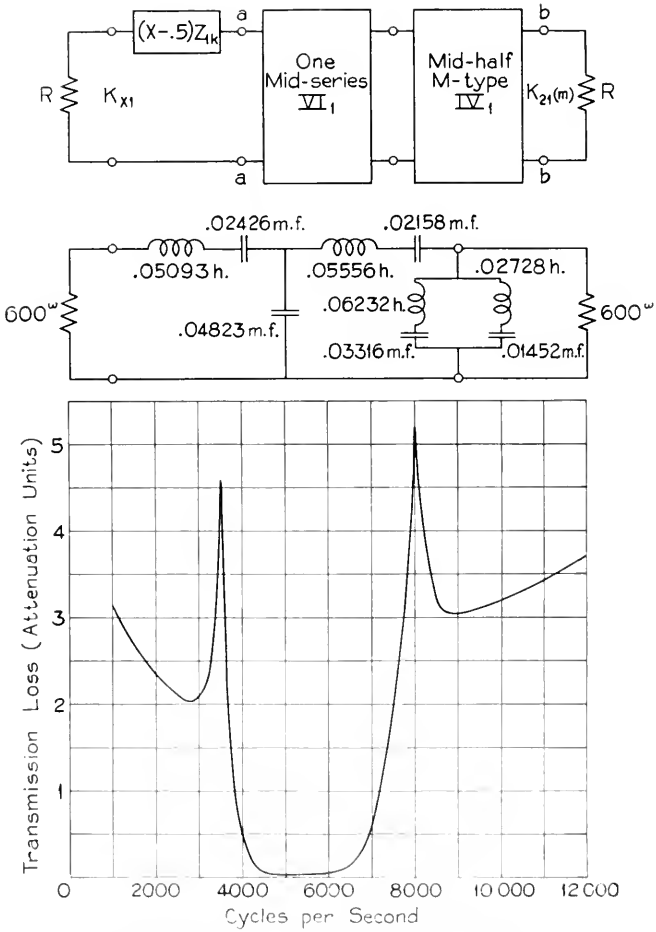


Fig. 4—Transmission Loss of Composite Band Pass Wave-Filter

APPENDIX

Derivation of Linear Transducer Formulae

The formula used in the text for a dissymmetrical composite wave-filter structure contains the image parameters [7] and is a special case of a general formula which is applicable to any linear transducer, active or passive. This general formula is derived here together with other useful ones.

A linear transducer will be defined as an electrical network which has two input and two output terminals and a structure such that so far as these terminals are concerned the currents are linear functions of the potential differences and therefore the principle of superposition holds. The structure may contain sources as well as sinks of energy; that is, the transducer may be active or passive. In the most general case, that of an active dissymmetrical linear transducer, four independent parameters are necessary to specify its electrical properties. Two sets of such parameters will be considered in deriving corresponding formulae, the image parameters and the recurrent parameters.

1. Image Parameters

1. *General Linear Transducer.* The parameters in this case are defined with reference to the single transducer in Fig. 3. Let the terminal impedances in this figure be so chosen that the impedances in the two directions from terminals a are equal, that is, the latter impedances are the "image" of each other, and at the same time a similar "image condition" holds with reference to terminals b. With the transducer so terminated, its *directional transfer constants* are here defined as $T_{ab} = \log_e (I_a / I_b)$ when transmitting from terminals a to terminals b, and $T_{ba} = \log_e (I_b / I_a)$ when transmitting from terminals b to terminals a. The *image impedance* W_a of the transducer is the impedance across terminals a in either direction, and the *image impedance* W_b is similarly defined at terminals b. In general, T_{ab} and T_{ba} are different, as are also W_a and W_b.

The transducer is now to be terminated by the general impedances Z_1 and Z_2 with an electromotive force E_1 applied in series with Z_1.

[7] The relations among five other distinct sets of parameters for a transducer such as a passive one which can be specified by three complex parameters were given by G. A. Campbell in Cisoidal Oscillations, Trans. A. I. E. E., Vol. XXX, Part II, Table I, p. 885, 1911. The different sets correspond to the four normal networks designated as the T, the Π, the transformer, and the artificial line, and to the simple circuit one-point and two-point impedances. A sixth set, one point and two-point admittances, was used in Appendix I of my paper in the B. S. T. J., Jan. 1923.

It is desired to obtain, among others, expressions for the sending end and receiving end currents I_a and I_b which contain the image parameters.

Each terminal impedance will be considered as equivalent to the image impedance at that end plus another impedance whose potential drop is to be replaced in the usual manner by an equal opposing electromotive force. In effect this equivalent electromotive force substitution reduces the system to one in which the transducer is terminated by its image impedances and in which determinate electromotive forces are acting at *both* ends. From this viewpoint, the total effective electromotive forces acting at the ends a and b of the transducer terminated by its image impedances W_a and W_b are, respectively,

$$E_a+(W_a-Z_a)I_a,$$

and

$$(W_b-Z_b)I_b. \tag{49}$$

Superposing the currents due to these electromotive forces at both ends we may write the current expressions immediately from the definitions of the parameters involved.

Thus

$$I_a=\frac{E_a+(W_a-Z_a)I_a}{2W_a}+\frac{(W_b-Z_b)I_b}{2W_b}e^{-T_{ba}},$$

and

$$I_b=\frac{E_a+(W_a-Z_a)I_a}{2W_a}e^{-T_{ab}}+\frac{(W_b-Z_b)I_b}{2W_b}. \tag{50}$$

Their solution gives the explicit formulae [10]

$$I_a=\frac{E_a}{W_a+Z_a}\frac{(1+r_be^{-(T_{ab}+T_{ba})})}{(1-r_ar_be^{-(T_{ab}+T_{ba})})},$$

and

$$I_b=\frac{E_a}{W_a+Z_a}\frac{(1+r_b)e^{-T_{ab}}}{(1-r_ar_be^{-(T_{ab}+T_{ba})})}, \tag{51}$$

where r_a and r_b, the current reflection coefficients at terminals a and b, are

$$r_a=\frac{W_a-Z_a}{W_a+Z_a},$$

and

$$r_b=\frac{W_b-Z_b}{W_b+Z_b}.$$

[10] These formulæ may also be derived synthetically by the current reflection method

Although the transducer has four independent parameters, it will be seen that the sending end current involves but three effective transducer parameters, the sum $(T_{ab}+T_{ba})$, W_a, and W_b. As a result, the four one-point impedance measurements which can be made upon the transducer itself, the open-circuit and short-circuit driving-point impedances at both ends, must have a relation between them. Let X_a and Y_a denote the driving-point impedances across terminals a when terminals b are open-circuited and short-circuited, respectively. Then if in (51) $Z_a=0$ and terminals b are open-circuited by putting $Z_b=\infty$, the impedance at terminals a, the *open-circuit impedance*, is

$$X_a=\frac{E_a}{I_a}=W_a\coth\tfrac{1}{2}(T_{ab}+T_{ba}). \tag{52}$$

Similarly for the *short-circuit impedance*, when $Z_a=0$ and $Z_b=0$,

$$Y_a=W_a\tanh\tfrac{1}{2}(T_{ab}+T_{ba}). \tag{53}$$

For the other end we get by interchanging subscripts

$$X_b=W_b\coth\tfrac{1}{2}(T_{ab}+T_{ba}), \tag{54}$$

and

$$Y_b=W_b\tanh\tfrac{1}{2}(T_{ab}+T_{ba}). \tag{55}$$

These give the necessary relation as

$$\frac{X_a}{Y_a}=\frac{X_b}{Y_b}. \tag{56}$$

Hence, *in the most general linear transducer the ratio of the open-circuit to short-circuit impedances at one end is equal to the corresponding ratio at the other end.*

Other important derived formulae are

$$T_{ab}+T_{ba}=2\tanh^{-1}\sqrt{\frac{Y_a}{X_a}}=2\tanh^{-1}\sqrt{\frac{Y_b}{X_b}}, \tag{57}$$

$$W_a=\sqrt{X_aY_a}, \tag{58}$$

$$W_b=\sqrt{X_bY_b}, \tag{59}$$

$$W_a-W_b=\sqrt{(X_a-X_b)(Y_a-Y_b)}, \tag{60}$$

and

$$W_aW_b=X_aY_b=X_bY_a. \tag{61}$$

Thus the open-circuit and short-circuit impedance measurements determine the *sum* of the directional transfer constants and both of the image impedances.

To obtain the separate values of T_{ab} and T_{ba}, it is necessary to make at least one two-point measurement, as seen from the formula for I_b which contains four distinct transducer parameters. For example, to find T_{ab} perhaps the simplest method is to terminate with the image impedance at terminals b, whence

$$T_{ab} = \log_e(I_a \sqrt{I_b}), \text{ where } Z_b = W_b. \tag{62}$$

The constant T_{ba} is the difference between the sum $(T_{ab}+T_{ba})$, obtained from (57), and T_{ab}; it may also be determined by a two-point measurement similar to the above for transmission in the opposite direction.

For some purposes it is convenient to have formulae involving the potential differences V_a and V_b across the two pairs of terminals, rather than the terminal impedances Z_a and Z_b and the series applied electromotive force E_a. Such formulae in combination with the above can be used to advantage in determining the currents and potential differences at points within a composite transducer. They are derived readily by making the substitutions in the above,

$$Z_a = \frac{E_a - V_a}{I_a},$$

and (63)

$$Z_b = \frac{V_b}{I_b}.$$

For current transmission from terminals a to terminals b

$$I_a = I_b\left(\frac{e^{T_{ab}} + e^{-T_{ba}}}{2}\right) + \frac{V_b}{W_b}\left(\frac{e^{T_{ab}} - e^{-T_{ba}}}{2}\right),$$

and (64)

$$V_a = V_b\frac{W_a}{W_b}\left(\frac{e^{T_{ab}} + e^{-T_{ba}}}{2}\right) + I_b W_a\left(\frac{e^{T_{ab}} - e^{-T_{ba}}}{2}\right).$$

Also,

$$I_b = I_a\left(\frac{e^{T_{ba}} + e^{-T_{ab}}}{2}\right) - \frac{V_a}{W_a}\left(\frac{e^{T_{ba}} - e^{-T_{ab}}}{2}\right),$$

and (65)

$$V_b = V_a\frac{W_b}{W_a}\left(\frac{e^{T_{ba}} + e^{-T_{ab}}}{2}\right) - I_a W_b\left(\frac{e^{T_{ba}} - e^{-T_{ab}}}{2}\right).$$

Interchanging the subscripts and changing the signs of the currents in (64) will also lead to (65).

2. *Passive Linear Transducer.* Since the reciprocal theorem holds here one relation exists between the four parameters leaving three

independent ones. This relation is given directly by the theorem in the case where $Z_1 = W_1$ and $Z_2 = W_2$, the equivalent transfer currents being

$$\frac{e^{I_a} - e^{-I_a}}{2W_1}, \quad \frac{2W_2}{e^{I_a} - e^{-I_a}} \tag{66}$$

Although any three of these parameters might be assumed as independent, it is convenient to take as the independent parameters I, W_1, and W_2, where

$$I = D + iS = \tfrac{1}{2}(I_{ab} + I_{ba}) \tag{67}$$

is thus defined for the *passive transducer* as the *transfer constant*. *The transfer constant is the arithmetic mean of the two directional transfer constants.* The real and imaginary parts of I, namely D and S, will be called the *diminution constant* and the *angular constant* to distinguish them from the attenuation constant and the phase constant of the ordinary propagation constant to which they reduce in the case of a symmetrical transducer. Then these parameters are given by the formulae

$$I = \tanh^{-1}\sqrt{\frac{Y_a}{Y_b}} = \tanh^{-1}\sqrt{\frac{Y_b}{Y_a}}$$

$$W_1 = \sqrt{Y_a Y_b}, \tag{68}$$

and

$$W_2 = \sqrt{Y_a Y_b},$$

and are completely determined by the open-circuit and short-circuit driving-point impedances.

With these parameters the current formulae become

$$I_1 = \frac{E_a}{W_1 + Z_a}\,\frac{(1 + r_{ba}e^{-2I})}{(1 - r_a r_{ba}e^{-2I})},$$

and

$$I_2 = \frac{E_a}{W_1 + Z_a}\sqrt{\frac{W_1}{W_2}}\,\frac{(1 + r_{ba})e^{-I}}{(1 - r_a r_{ba}e^{-2I})}, \tag{69}$$

$$= \frac{2E_a\sqrt{W_a W_b}\,e^{-I}}{(W_1 + Z_a)(W_2 + Z_b)(1 - r_a r_{ba}e^{-2I})},$$

the latter being the one used in the text. Other forms are

$$I_1 = \frac{E_a Z_b \sinh I + W_2 \cosh I}{(W_1 W_2 + Z_a Z_b)\sinh I + (W_1 Z_b + W_2 Z_a)\cosh I},$$

and

$$I_2 = \frac{E_a\sqrt{W_2 W_1}}{(W_1 W_2 + Z_1 Z_2)\sinh I + (W_1 Z_2 + W_2 Z_1)\cosh I} \tag{70}$$

Introducing the potential differences, for current transmission from terminals a to terminals b

$$I_a = I_b\sqrt{\frac{W_b}{W_a}}\cosh T + \frac{V_b}{\sqrt{W_a W_b}}\sinh T,$$

and (71)

$$V_a = V_b\sqrt{\frac{W_a}{W_b}}\cosh T + I_b\sqrt{W_a W_b}\sinh T.$$

Also,

$$I_b = I_a\sqrt{\frac{W_a}{W_b}}\cosh T - \frac{V_a}{\sqrt{W_a W_b}}\sinh T,$$

and (72)

$$V_b = V_a\sqrt{\frac{W_b}{W_a}}\cosh T - I_a\sqrt{W_a W_b}\sinh T.$$

II. Recurrent Parameters

1. *General Linear Transducer.* Here four parameters[11] of the transducer in Fig. 3 are defined in terms of its properties when it is one section of an infinite recurrent structure which is made up of identical sections, similarly oriented. With such terminal conditions for the transducer, its *directional propagation constants* are defined as follows: $\Gamma_{ab} = \log_e (I_a, I_b)$ when transmitting from terminals a to terminals b, and $\Gamma_{ba} = \log_e (I_b/I_a)$ when transmitting from terminals b to terminals a. The *characteristic impedance* K_a is the impedance across terminals a in the direction from a to b, and the *characteristic impedance* K_b is similarly defined for the impedance across terminals b in the opposite direction.

Terminating the transducer by the general impedances Z_a and Z_b and applying an electromotive force E_a in series with Z_a, the current formulae containing the recurrent parameters may be derived in a manner analogous to that used with the image parameters. In this case the total effective electromotive forces acting at the ends a and b of the transducer terminated by its characteristic impedances K_b and K_a are, respectively,

$$E_a + (K_b - Z_a) I_a,$$

and (73)

$$(K_a - Z_b) I_b.$$

[11] These parameters may also be designated in the general case as those of a generalized artificial line.

Hence,

$$I_a = \frac{E_a}{K_a + Z_a} \frac{(1 + \rho_b e^{-(\Gamma_{ab} + \Gamma_{ba})})}{(1 - \rho_a \rho_b e^{-(\Gamma_{ab} + \Gamma_{ba})})},$$

and (74)

$$I_b = \frac{E_a}{K_a + Z_a} \frac{(1 + \rho_b) e^{-\Gamma_{ab}}}{(1 - \rho_a \rho_b e^{-(\Gamma_{ab} + \Gamma_{ba})})},$$

where the current reflection coefficients at terminals a and b are

$$\rho_a = \frac{K_b - Z_a}{K_a + Z_a},$$

and

$$\rho_b = \frac{K_a - Z_b}{K_b + Z_b}.$$

Introducing the open-circuit and short-circuit driving-point impedances X_a, X_b and Y_a, Y_b of the transducer it follows that

$$\Gamma_{ab} + \Gamma_{ba} = 2 \tanh^{-1} \sqrt{\frac{(X_a - X_b)^2 + 2(X_a \bar{Y}_b + \bar{X}_b Y_a)}{X_a \bar{Y}_a + X_b}}, \tag{75}$$

$$\left.\begin{matrix} K_a \\ K_b \end{matrix}\right\} = \frac{1}{2}\left[\sqrt{(X_a - X_b)^2 + 2(X_a Y_b + \bar{X}_b Y_a)} \pm (X_a - X_b)\right], \tag{76}$$

$$K_a - K_b = X_a - X_b, \tag{77}$$

and

$$K_a K_b = X_a Y_b = X_b Y_a. \tag{78}$$

Any three of these measured impedances are sufficient, because of relation (56), to obtain the *sum* $(\Gamma_{ab} + \Gamma_{ba})$, K_a, and K_b.

A directional propagation constant may be obtained separately from one two-point measurement; thus

$$\Gamma_{ab} = \log_e (I_a / I_b), \text{ where } Z_b = K_a. \tag{79}$$

The current and potential difference at one pair of terminals in terms of those at the other are given by the following.

For current transmission from terminals a to terminals b

$$I_a = I_b\left(\frac{K_b e^{\Gamma_{ab}} + K_a e^{-\Gamma_{ba}}}{K_a + K_b}\right) + \frac{V_b}{K_a + K_b}(e^{\Gamma_{ab}} - e^{-\Gamma_{ba}}),$$

and (80)

$$V_a = V_b\left(\frac{K_a e^{\Gamma_{ab}} + K_b e^{-\Gamma_{ba}}}{K_a + K_b}\right) + I_b \frac{K_a K_b}{K_a + K_b}(e^{\Gamma_{ab}} - e^{-\Gamma_{ba}}).$$

Also,

$$I_b = I_a\left(\frac{K_a e^{V_{ba}} + K_b e^{-V_{ab}}}{K_a + K_b}\right) - \frac{V_a}{K_a + K_b}\left(e^{V_{ba}} - e^{-V_{ab}}\right),$$

and (81)

$$V_b = V_a\left(\frac{K_b e^{V_{ba}} + K_a e^{-V_{ab}}}{K_a + K_b}\right) - I_a\frac{K_a K_b}{K_a + K_b}\left(e^{V_{ba}} - e^{-V_{ab}}\right).$$

2. *Passive Linear Transducer.* Because of the reciprocal theorem the directional propagation constants become equal giving a single *propagation constant*,

$$\Gamma = A + iB = \Gamma_{ab} = \Gamma_{ba}, \tag{82}$$

which is obtainable from the general formula (75). Here A is the *attenuation constant* and B is the *phase constant*.

The current formulae become

$$I_a = \frac{E_a}{K_a + Z_a}\frac{(1 + \rho_b e^{-2\Gamma})}{(1 - \rho_a\rho_b e^{-2\Gamma})},$$

and (83)

$$I_b = \frac{E_a}{K_a + Z_a}\frac{(1 + \rho_b)e^{-\Gamma}}{(1 - \rho_a\rho_b e^{-2\Gamma})}.$$

In the other form they are

$$I_a = \frac{E_a[(1 - K_a + K_b + 2Z_b)\sinh\Gamma + (K_a + K_b)\cosh\Gamma]}{[(2(K_a K_b + Z_a Z_b) - (K_a - K_b)(Z_a - Z_b))\sinh\Gamma + (K_a + K_b)(Z_a + Z_b)\cosh\Gamma]},$$

and (84)

$$I_b = \frac{E_a(K_a + K_b)}{[(2(K_a K_b + Z_a Z_b) - (K_a - K_b)(Z_a - Z_b))\sinh\Gamma + (K_a + K_b)(Z_a + Z_b)\cosh\Gamma]}.$$

Introducing the terminal potential differences, when transmitting from terminals a to terminals b

$$I_a = I_b\left(\cosh\Gamma - \frac{K_a - K_b}{K_a + K_b}\sinh\Gamma\right) + \frac{V_b}{K_a + K_b}2\sinh\Gamma,$$

and (85)

$$V_a = V_b\left(\cosh\Gamma + \frac{K_a - K_b}{K_a + K_b}\sinh\Gamma\right) + I_b\frac{K_a K_b}{K_a + K_b}2\sinh\Gamma;$$

and at the other terminals

$$I = I_1\left(\cosh\Gamma + \frac{K_a - K_b}{K_a + K_b}\sinh\Gamma\right) \qquad \frac{V_2}{K_a + K_b}2\sinh\Gamma,$$

and (86)

$$V = V_1\left(\cosh\Gamma + \frac{K_a - K_b}{K_a + K_b}\sinh\Gamma\right) \qquad I_2\frac{K_aK_b}{K_a + K_b}2\sinh\Gamma.$$

Comparison shows that the general formulae for the currents I_1 and I_2 given by (51) and (74) in terms of the two sets of parameters are of the same functional form involving their respective reflection coefficients; the latter are of slightly different functional forms. This similarity is what one expects when deriving the formulae synthetically by the current reflection method.

In all cases by (61) and (78)

$$W_aW_b = K_aK_b. \tag{87}$$

The sum $(Z_a + Z_b)$, W_a, and W_b of any transducer are obviously also equal to the propagation constant and respective characteristic impedances of the two symmetrical transducers which can individually be formed with two such identical transducers.

If $Z_{ab} = Z_{ba}$, the reciprocal theorem holds only when $W_a = W_b$, for which case the transducer is symmetrical. On the other hand if $\Gamma_{ab} = \Gamma_{ba}$, this theorem holds irrespective of the values of K_a and K_b. In each of these cases which satisfies the theorem the transducer may be active or passive.

In an electrically symmetrical transducer, whether active or passive, two parameters specify its properties
where

$$Z_{ab} = Z_{ba} = \Gamma_{ab} = \Gamma_{ba},$$

and (88)

$$W_a = W_b = K_a = K_b,$$

in which case the corresponding formulae are identical in the parameters. Structural symmetry is not necessary here as may be seen, for example, in the case of a composite wave-filter made up of different mid-series sections whose characteristic impedances are equivalent.

In a passive dissymmetrical transducer the formulae containing hyperbolic functions are of simpler form with the parameters I, W_a, and W_b than with the parameters Γ, K_a, and K_b. The image parameter formulae are readily applicable where the transducer is made up of parts whose image impedances at the junctions are equivalent,

as in the present case of a composite wave-filter. Simple relations exist here between these parameters of the transducer and of its parts, as shown in the text, which is not true with the other parameters. The recurrent parameter formulae, on the other hand, apply more naturally when dealing with a succession of identical dissymmetrical sections, or of different dissymmetrical sections whose characteristic impedances in one direction are equivalent, in which cases the propagation constant of the transducer is equal to the sum of the progagation constants of the parts. In conclusion, it is seen that the set of parameters most suitable for use in any case depends upon the particular structure of the transducer.

Some Contemporary Advances in Physics—V

By KARL K. DARROW

Electricity in Solids

IN considering such topics as the flow of electricity through solids and the outflow of electricity across their boundaries, we have to forego the assistance of the great system of laws, models, and word-pictures which constitutes the contemporary theory of the structure of the atom. This imposing and truly powerful theory, which nowadays seems to bulk larger than all of the rest of physics, is after all limited to certain restricted fields; it deals successfully with particular properties of isolated atoms, and also with certain qualities of atoms which seem to be localized in their inner regions; but it avails little or nothing in the study of the behavior of liquids and solids. Much of the present-day theory of electrical conduction in solids is based only on the very simplest assumptions as to the nature of the atoms of which they are built, some would even remain valid under the old-fashioned ideas of continuous electrical fluids; and profoundly as we may believe that solids are built of atoms resembling Bohr's famous model, it is highly doubtful whether that model has ever helped to interpret a single one of the phenomena of conduction or done more than to provide a new language for old ideas.

We have first to make the distinction between the substances in which atoms migrate along the path of the flowing current and apparently carry the moving charge, and the substances in which the atoms stand still while the current flows past them. It is universally conceded that elements, and likewise the alloys of metals and a number of solid compounds, belong to the latter class; whatever it is that carries the current flows through and past the substance, leaving it at the end as it was at the beginning. Weber said in 1858, "In the metals there are electrically-charged particles as well as atoms; some of the former are freely mobile and others vibrate about the atoms; they are the cause of the conduction of electricity and of heat, and of magnetic phenomena as well." Considering that in Weber's day electricity had never been observed apart from ponderable matter and electrons were unknown, this is entitled to rank as a daring anticipation.

Next we have to distinguish between conduction by metals and conduction by non-metallic elements. Strictly we should begin by defining a "metal"; but this task had better be left to the chemists, as being really their affair; and they have found it no easy affair to

PERIODIC CLASSIFICATION OF THE ELEMENTS[1]

	I	II	III	IV	V	VI	VII	VIII			O
1	1 H										2 He
2	3 Li	4 Be	5 B	6 C	7 N	8 O	9 F				10 Ne
3	11 Na	12 Mg	13 Al	14 Si	15 P	16 S	17 Cl				18 A
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	
	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br				36 Kr
5	37 Rb	38 Sr	39 Yt	40 Zr	41 Nb	42 Mo	43	44 Ru	45 Rh	46 Pd	
	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I				54 Xe
6	55 Cs	56 Ba	Rare Earths	72 Hf	73 Ta	74 W	75 —	76 Os	77 Ir	78 Pt	
	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 —				86 Nt
7	87 —	88 Ra	89 Ac	90 Th	91 Pa	92 U					

[1] The atomic numbers are given in front of the abbreviations of the elements.

set up a definition by which every element can be confidently assigned to one class or to the other. In fact there is a tendency to begin by defining metallic conduction, and then define metals as the elements which display it! The difficulty, as usual, is to make the definition sharp enough to decide a few intermediate or transitional cases. Anyone even slightly acquainted with chemistry or physics would instantly recognize as metals the elements in the first column of the Periodic Table, and those at the bottom of the table in all the columns; and as non-metals, with the same ease, the elements in the topmost row of the table and down the right-hand side. The first element of every column after the first two is non-metallic, and the non-metallic character advances farther and farther down the columns as one proceeds across the Table from left to right. One might say that the elements which are not metals occupy the north-east sector of the Table, and the debatable ones cross in a diagonal band from northwest to southeast. The elements which are gases under the usual circumstances of temperature and pressure are extreme instances of non-metals; but some of the definitely non-metallic elements, and all of the debatable ones, are solid or liquid under the usual conditions.

Very little could be said about the elements which under ordinary conditions are gases, for very little is known about the manner in which they conduct electricity when liquefied or frozen. Probably the reason is that the experimental conditions would be unusually difficult, and the substances probably very bad conductors; it is not easy to imagine solid hydrogen moulded into a cylinder, drawn into a wire, clamped or sealed between electrodes, or filled into a sheath less conductive than the hydrogen itself. The difficulties may not be insuperable; but they have not been generally overcome.

As for the solid elements which are definitely not metals, or which belong to the debatable group, there is an abundance of data in print, and yet not nearly so much as we need. In general their resistances are tremendously greater than the resistances of metals; "tremendously" for once is not an extravagant word, for the conductivities of the elements are spread over a sweeping range of orders of magnitude which few if any other qualities of theirs can rival. The mass of the heaviest known atom differs from the mass of the lightest only by a factor of 240; the densities of the solidified elements, their compressibilities, their other mechanical and thermal properties range over not more than one or two, at the most three orders of magnitude; even the energy required to extract the innermost electron of an atom rises by a factor of only 10 in passing from the first to

the last element of the series; but the conductivity of silver stands to the conductivity of sulphur in the ratio 10^{21}. The distance from the sun to the nearest star is some 10^{18} cm.; we see that a sheet of sulphur a thousandth of an inch thick would offer more of an obstacle to the passage of electricity than a cable of silver of the same diameter, extending from the earth to Alpha Centauri. The variations of conducting-power from element to element are thus as fantastically great as the variations in scale from the world of common life to the world of interstellar spaces. The conductivities of the metals, however, are confined within a narrow fraction of this range; it is between the metals and the non-metals, and between one non-metallic element and another, that the leaps are surprisingly great.

In general, too, the resistance of a non-metallic element decreases as its temperature is raised; the curve of resistance versus temperature (I shall often call it *characteristic*, henceforward) slants downward, the derivative and the temperature-coefficient of resistance are negative. Near room-temperature this is the usual behavior, but not always over the entire accessible range; of some elements it is observed that the resistance declines less rapidly as the temperature is raised, the curve is concave upward; eventually the decline ceases, the resistance passes through a minimum value at a certain characteristic temperature, and thereafter increases with the temperature as the resistances of metals do. At least one element of the debatable class (germanium) exhibits a characteristic curve that slants upward instead of downward at room-temperatures; but when the curve is followed towards lower temperatures, it too is found to be concave upward with a minimum of resistance below $-100°$ C. This suggests that for all of the non-metals the resistance-temperature curve may be a loop bulging downward, with a minimum at a certain temperature that varies from element to element; on this generalization one of the contemporary theories is founded.

These rules can be illustrated by mentioning briefly the behavior of the non-metallic elements one by one. Beginning at the foot of the procession of elements, we pass over hydrogen (no data), lithium and beryllium (metals), and commence with boron. *Boron* has a very high resistance at room temperature, which drops a hundredfold when it is heated to 180° C. and ten-million-fold when it is raised to a red heat. On *carbon* a tremendous amount of work has been done, which unfortunately largely goes to show that the word "carbon" usually signifies a framework of carbon atoms packed with occluded gases, organic compounds, and impurities of divers kinds, which no known mode of treatment avails to expel entirely, although almost

anything which is done to the substance alters its constitution enough to affect its resistance. (We shall later see that the situation with many of the metals is almost as bad.) Most of the experiments reveal a steady decline of resistance as the temperature is raised, whether the sample used be amorphous or crystalline (graphitic) and whatever its history; but Noyes recently traced several very concordant curves for several samples of graphite (all however of the same provenience) showing a minimum of resistance near 800° C. Diamonds have exceedingly high resistances, which fall when they are heated.

Passing over four gases and three metals, we come next to *silicon*; the curve traced by Koenigsberger shows the resistance descending as the temperature is increased, until at a certain critical temperature it leaps sharply upward; from the new high value it descends again as the silicon is further warmed, only to make a second upward jump; from this second maximum it drops steadily away, at least as far as the highest temperature attained in the experiment. This illustrates another perplexing property of some elements; they have several distinct "allotropic" forms, each of them more or less stable over a distinct range of temperature which may or may not overlap with the ranges of the others; each must be regarded, so far as its conducting-power is concerned, as a distinct element. In some instances the several forms of an element are vividly contrasted in appearance and in general behavior; such is the case with *phosphorus*, all of the forms of which have high resistances, but little is known about their trends with temperature. In other cases the anomalous changes of temperature with resistance are not accompanied by other striking changes; and there is a tendency to explain any deviation from an expected trend—such as, for example, a maximum in a resistance-temperature curve—by saying that the substance is gradually changing from one form into another.

Sulphur is the extreme case of high resistance. I know of no data for *scandium*, which is to be regretted, as there is some reason from general atomic theory for supposing that this element stands at a turning-point of the Periodic Table. *Titanium*, like silicon, has several modifications, in some of which the characteristic rises while in others it descends. *Germanium* has been studied lately by Bidwell; it is the element mentioned above which displays a minimum of resistance at −116° C. *Arsenic* resembles the metals. *Selenium* in the dark has an extremely high resistance; its character when illuminated is too much of a subject to be discussed in this place. *Zirconium* was found, at least by one observer, to display a minimum

of resistance at 70° C., though in conductivity it compares favorably with the accepted metals. *Antimony*, although ranked among the metals, is usually to be found among the exceptions to any rules laid down for them; the same can be said of *bismuth*. *Tellurium* is an outstanding instance of an element with two modifications, and a sample taken at random is likely to be a mixture of them in unpredictable proportions, which change when it is heated; the characteristics are correspondingly crooked, and rarely agree. *Iodine* has a very high resistance.

Comparing the metals as a group with the non-metals, the first striking rule is that their conductivities are much higher and rather close together; from silver (the most conductive of all substances at room-temperature), to bismuth, the most resistant of the elements commonly accepted as metals, the conducting-power descends in the relatively small ratio of 75 to 1. The next and familiar rule is, that increasing temperature and increasing resistance always go together; the characteristic always slants upward to the right, the derivative and the temperature-coefficient of resistance are positive. It is customary to say that the resistance is always approximately proportional to the temperature, and that the temperature-coefficient of resistance always has approximately the one universal value, which is the value of the temperature-coefficient of volume of an ideal gas at constant pressure (or its temperature-coefficient of pressure at constant volume). That is to say, when the temperature of a piece of metal is increased by a given amount, its resistance increases approximately in the same proportion as would the pressure of a fixed quantity of an ideal gas, enclosed in a non-expanding container and raised from the same initial to the same final temperature as the metal. Were these statements literally true, all the resistance-temperature curves for metals would be straight lines intersecting the axis of temperatures at absolute zero. But the second statement cannot even be considered a good approximation, unless one is willing to confer the title "good approximation" on a numerical value .00365 which is expected to agree with a set of observed values which ranges upwards to .0058 (potassium) and .0063 (iron). (I refrain from giving a lower limit for the range, for a reason which will presently be made clear.) Also the characteristic curves are not rigorously straight lines, although it is not unreasonable to call some of them *approximately* straight, when one considers how wide is the interval of temperature over which some of them have been traced. In some cases a quadratic term added to the linear expression, resulting in a formula $R = R_0 + at + bt^2$ is sufficient to express the data. Usually,

but not always, the extra coefficient b is positive; the characteristic is concave upward. "Usually but not always" is a phrase much in demand when one is laying down rules for conducting bodies. In this case metals of the platinum triad furnish the exceptions. In other instances cubic terms must be added to the formula, and in still others even these are inadequate. One of the longest characteristics ever traced, the one determined by Worthing and Forsythe for tungsten from 1100° to 3250° C., conforms to the equation $R = const.\ T^{1.2}$.

All these details about values of resistances and shapes of resistance-temperature curves are sedate and commonplace enough; but there is one quite extraordinary phenomenon in this field, one of the strange discontinuities which appear here and there in the theatre of nature and contribute more of dramatic interest to the spectacle than any amount of smooth correlations between continuous variables. Extensions of the characteristics downwards toward the absolute zero have to follow upon improvements in the art of producing and maintaining very low temperatures; and for the last twenty years the advances in this art have been made in the Cryogenic Laboratory of the University of Leyden, and there the curves have been extended downwards step by step as additional ranges of cold were made accessible. The temperatures down to 14° K. attained with liquefied hydrogen did not affect the resistances of metals in any very startling way, although the characteristics are generally more sharply curved there than at ordinary temperatures; but when with the aid of liquefied helium Kamerlingh Onnes penetrated to within five degrees of the absolute zero, something astonishing took place.

Kamerlingh Onnes had been experimenting with platinum wire, and he had found that over the interval of temperatures newly made available, the interval from 4.3° to 1.5° K. (a small range when measured in degrees, but a great one when considered in terms of the distance between its lower limit and the absolute zero) the resistance of the wire did not change. This he thought might mean that the proper resistance of the metal had become exceedingly small, leaving as the chief component of the observed resistance a term unaffected by temperature and due possibly to some such thing as discontinuities in the wire, for example between the platinum and bits of impurities mixed into it. To have a purer metal he replaced the platinum by repeatedly-distilled mercury. It was contained in a slender glass capillary tube, forming so fine a filament that the resistance at room-temperature was rather considerable; in one specified instance, 173 ohms. When he lowered this filament of mercury to the temperature

of frozen helium, at a certain point the resistance suddenly vanished. Literally it vanished; the word is justified, for the value to which it had dropped was, if not truly zero, at all events not so much as one five-billionth of its value at room-temperature, and not so much as one ten-millionth of its value just before, at about 4.1° K., it suddenly disappeared. The mercury had altogether lost what had always seemed to be as inseparable a quality of matter as its inertia or its weight.

A few other elements were later found to share this property; tin, of which the resistance vanishes at 3.78°; lead, having its threshold-temperature at 7.2°; thallium, at 2.3°. Three of these four are consecutive in the procession of elements. Other elements were definitely found not to become "supra-conductive" within the accessible range: gold, cadmium, platinum, copper and iron. In the vicinity of the absolute zero each of these metals has a constant resistance independent of temperature. This as I mentioned was interpreted to mean that these metals, or at least these samples, behaved thus because they were impure—that impurities prevented the vanishing of resistance—but since mercury contaminated intentionally with gold or with cadmium was found to become supra-conductive, and tin amalgam likewise, it has become necessary to save this interpretation, if at all, by assuming that in the five specified metals the impurities coalesce with the metal in some particular way. It is interesting to note that the threshold-temperature of tin amalgam lies above that of either of its components—at 4.29° K., to be compared with the 4.1° of mercury and the 3.78° of tin. These thresholds are not entirely independent of circumstances; they diminish when a large current-density is used, and also when a magnetic field is applied, possibly from the same reason in both cases.

A number of fantastic things could happen in a world from which electrical resistance had vanished, and one of them was actually realized by Kamerlingh Onnes within the compass of his helium-cooled chamber, when a current of three hundred and twenty amperes flowed for half-an-hour around and around a leaden ring with no applied E.M.F. whatever to maintain it, and did not lose as much as one one-hundredth of its initial strength. In another experiment a current of forty-nine amperes flowed for an hour around a coil of lead wire of a thousand turns, wound upon a brass tube, and did not lose quite one per cent. of the intensity with which it had been started by removing a magnet of which the field had interlaced the coil. At this rate it would have taken over four days for the current to drop to the 1 e-th part of its initial value, if the coil could have been

kept cold so long. This corresponds to a resistance lower than 3.10^{-7} ohms; the resistance of the coil at room-temperature was 731 ohms. Few discoveries in physics can have been so exciting as this one, and further news from Leyden is awaited with keen anticipation. Until the present liquefied helium has been made nowhere else, but from now on the process will be carried on at Toronto also.

Pressure affects the resistance of a metal much less than temperature; that is to say, doubling the hydrostatic pressure upon a metal makes no perceptible difference with its resistance if the initial pressure is one atmosphere or less, and usually alters it only by a few per cent. if the intial pressure amounts to thousands of atmosphere. The art of applying enormous pressures under controllable conditions has been developed furthest by Bridgman in the Physical Laboratory of Harvard University, which through his work holds the same unique rank in high-pressure investigations as Kamerlingh Onnes' laboratory at Leyden in low-temperature research. The highest pressure which Bridgman has applied to metals during resistance-measurements exceeds 12,000 *kg cm*$^{-2}$, which amounts practically to twelve thousand atmospheres. No one has ever applied temperatures twelve thousand times as great as room-temperature, nor even four thousand times as great as the lowest accessible temperature; but when the pressure is altered in this enormous ratio the resistance changes only by a few per cent. The volume likewise changes by only a small fraction, which rather suggests that it is the change in closeness of packing of atoms rather than the creation of intense internal stresses which is responsible for the change in conductivity; however, there is no close correlation between relative change in volume and relative change in resistance; sometimes the two are of opposite signs. Usually, but not always, the conductivity increases with the pressure; as if squeezing the atoms together facilitated the flow of electricity across the metal. The rule applies to thirty-five elements, distributed as follows in the Periodic Table: in the first column, 11 Na, 19 K, 29 Cu, 47 Ag, 79 Au; second column, 12 Mg, 30 Zn, 48 Cd, 80 Hg; third, 13 Al, 31 Ga, 49 In, 81 Tl; fourth, 6 C, 22 Ti, 40 Zr, 50 Sn, 82 Pb; fifth, 15 P, 33 As, 73 Ta; sixth, 42 Mo, 52 Te, 74 W, 92 U; seventh, 53 I; eighth, 26 Fe, 27 Co, 28 Ni, 45 Rh, 46 Pd, 77 Ir, 78 Pt; rare earths, 57 La, 60 Nd. Several of the non-metallic elements are found in the list. The exceptions are the five curiously assorted metals 3 lithium, 20 calcium, 38 strontium, 51 antimony, 83 bismuth five elements distributed over three columns of the Periodic Table, each of which contains several other elements which conform to the rule. One modification of 55 caesium belongs under the rule, another among

the exceptions. This illustrates how the behavior of metals in conducting electricity is liable to cut across the classification of the Periodic System, which controls nearly all of the properties of elements except those that vary uniformly from one element to the next all along the series.

As for the magnitude of the effect, the resistances of most metals are decreased through less than 10% by applying a pressure of ten thousand atmospheres, some only through one or two per cent.; but the decrease is 40% for sodium, 70% for potassium, 70% also for the "debatable" element tellurium, and 97% for black phosphorus; bismuth gains about 25% in resistance and antimony about 10%. The curves representing resistance as function of pressure are somewhat curved, but not greatly so; however the curvature frequently varies along the curve to such an extent that a two-constant formula is not sufficient to express the data. It is an interesting fact that the percentage by which a given pressure changes the resistance of a metal is approximately independent of its temperature, and consequently the percentage by which a given rise in temperature changes the resistance is approximately independent of the pressure; so that the combined effects of a pressure-change Δp and a temperature-change ΔT on a metal change its resistance from R_0 to $R_0(1+a\Delta p)(1+b\Delta T)$.

Tension, which is equivalent to negative pressure acting along a particular direction (there is no way of applying a negative hydrostatic pressure) results in lengthening the metal along one direction, shortening it along all directions perpendicular to that one, and dilating it as a whole. Most of the information about what it does to electrical resistance is owed to Bridgman. Usually, but not always, tension increases the resistance to current-flow along the direction of the stress. The exceptions are bismuth and strontium. Comparing the data about the effects of pressure and of tension, we see that Bi and Sr are exceptions to the common rules for both, while Li, Ca and Sb are exceptions to the usual rule for pressure but not to the usual rule for tension. This helps to show why it is so difficult to set up a thoroughly satisfactory theory of conduction in metals.

By melting a substance its density can be altered without altering either its temperature or its pressure; of course, the balance of interatomic forces is also altered in some mysterious but very potent way. Melting a solid usually brings about a decrease in density; the solid sinks in the liquid; but there are exceptions (bismuth, antimony, gallium). The conductivity always changes in the same sense as the density; hence for most metals the solid is more conductive than the

liquid, but bismuth, antimony, and gallium have greater resistances frozen than molten. This is one of the few rules in this field to which no exceptions have yet been discovered. The observed values of the ratio (resistance of liquid)/(resistance of solid), when tabulated and examined, show a tendency to cluster about values which are ratios of simple integers, such as 2:1, 4:3, 1:1. It would probably require a careful and expert analysis to show whether this tendency is more pronounced than a quite random distribution might reasonably be expected to display. Mercury has the highest ratio of all, 1:1.

Other agencies which are harder to measure or control may have distressingly great effects on the conductivity of a metal. The various metallurgical processes, annealing, cold-working and the rest, affect the resistance; sometimes the sign of the change can be explained by saying that the process has caused the many small crystals forming the metal to fuse into a few large ones, diminishing the resistance offered by the intercrystalline partitions; sometimes this explanation fails to work. Impurities may have a serious effect; for example Bridgman remarks of bismuth that "a fraction of a per cent. of lead or tin may change the temperature-coefficient from positive to negative and increase the specific resistance severalfold." Often impurities betray themselves by an abnormally low temperature-coefficient of the metal; this means that the absolute rate of increase is unusually small compared to the value of the resistance itself. This is so generally the case that a value of temperature-coefficient which (at 0° C.) is much below, say, .004 is usually taken to mean that the sample of metal under investigation is impure; and the "standard" values for individual metals set down in tables have often taken sudden jumps upward, when better-purified samples became available for measurements. For this reason I laid more stress, in a preceding paragraph, on the values which far exceed .00365 rather than the values which fall far below it. A metal contaminated by a small admixture of another metal may be regarded as the limiting case of an alloy. There is an enormous literature of the electrical behaviour of alloys, and some of the results can be extended to this limiting case. It is found, for example, that if two metals A and B form mixed crystals with one another, an alloy formed by mixing a small percentage or a fraction of one per cent of A into B, has a surprisingly greater resistance than B; and vice versa. The temperature-coefficient of the alloy is on the other hand much smaller than that of the metal, and may even be negative. Thus, although an alloy of this type may seem to be as thoroughgoing a metal as either of the pure elements of which it is made, it has a thoroughly anomalous

electrical behaviour; and the alloys as a whole, instead of assisting us to understand conduction in metals, contribute generously to the already abundant supply of difficulties. It remains to be seen whether the measurements upon single crystals of metals, which are being published at a steadily-increasing rate, are going to clarify the situation or increase the perplexity.

While I have left unmentioned a large number of the phenomena which a theory of conduction must be required to explain, the few which I have described will give quite an adequate basis for beginning a discussion of some of the extant theories. It must be conceded at once that the situation is bad. Perhaps there is some set of assumptions or of postulates by which the whole chaotic crowd of phenomena can be unified into a harmonious system; but if so, no one has yet formulated it. The theories, such as they are, may be divided into two groups: theories in which the electrons are supposed to move freely within the atoms and be stopped when they reach an interspace, and theories in which the electrons are assumed to move freely within the interspaces and be stopped when they collide with atoms. Those of the first kind start out with the advantage of being better adapted to the usual effect of pressure on resistance; most metals become more conductive when compressed, as if conduction were assisted by squeezing the atoms closer together. Still the oldest, the best-known, and the most highly elaborated of all the theories belongs to the second kind. This is the one formally known as the electron theory of metallic conduction, or more briefly as the electron theory of metals, and quite commonly as the "classical" theory of conduction (it does not take an idea so long to become "classical" in physics as it does in the arts). Founded by Riecke and by Drude in the closing years of the last century, it was developed by Lorentz and has since been worked over by Planck, Wien, Bohr, and other savants of the first eminence. Its popularity is largely due, I suspect, to the fact that it can be formulated with great if specious exactness: that is to say, as soon as a few definite assumptions are made (such as the simple, if unplausible, assumptions that the atoms are big elastic spheres and the electrons little ones), numerical consequences can be calculated with any degree of precision. In this respect most of the competing theories are sadly defective. Two or three of the numerical deductions made from simple auxiliary assumptions have agreed rather well with experimental data; and they have contributed to the feeling that there must be some kernel of truth in the mathematics, even if not in the physics of the thing, although it breaks down in so many other comparisons with experiment.

Fundamentally the theory is very simple, and has not been helped to any great extent by the more sophisticated mathematics which it emendators have introduced into it. What is observed in electrical conduction is this: when a potential-difference is established across a piece of metal, the electrons do not fall freely clear across it and emerge at the positive end with all the kinetic energy which the P.D. should have communicated to them; they ooze gradually through the metal, heating it as they go along and emerging with no unusual amount of energy, as if they had rubbed along through the metal like heavy particles dropping at constant speed through a gas. "Rubbing along" being a concept foreign to the atomic scale, we have to interpret that each electron falls freely through a small distance, collides with something to which it gives up the energy acquired from the field during its fall, falls again across another short distance, gives up its new quota in another collision, and so forth from side to side of the metal. Furthermore the energy which it gives up at each stoppage must find its way directly or indirectly into the heat of the metal, i.e., into thermal agitation of its atoms. Representing by T the time-interval between two consecutive collisions, by E the field-strength in the metal, by e and m the charge and mass of the electron, by U the average kinetic energy acquired by the electron from the field in its free fall between two collisions, we have

$$U = \tfrac{1}{2}(eET/m)^2 m. \tag{1}$$

If there are n electrons in unit cube of the metal, and each is stopped $1/T$ times in unit time, the rate at which heat appears in the unit cube is nU/T; but this rate is by definition the product of the conductivity σ by the square of the field-strength E, hence

$$\sigma = \tfrac{1}{2} ne^2 T/m \tag{2}$$

The same equation (2) can be reached, if one prefers to think of conductivity as the ratio of current-density to field-strength, by considering that during each free fall, the field augments the speed of each electron in the direction of the field-vector by the amount eET/m, which on the average is lost at the collision terminating the fall; so that the result is as if the field imprinted a constant drift-speed equal to $\tfrac{1}{2}eET/m$ upon all the electrons. Multiplying by ne to get the current-density and dividing by E to get the conductivity, we arrive again at (2).

Equation (2) is the fundamental equation of the electron theory of conduction, and indeed of most of the other theories. Let us begin by trying the supposition that the electrons are at rest until the field

is applied, and are brought to a full stop at each collision. Represent by l the average distance traversed between collisions. The proposed assumption leads to $I = \sqrt{2ml}/eE$. The conductivity therefore would depend on the field-strength, which would violate Ohm's law. Ohm's law being rigorously valid except under extreme conditions (Bridgman found the first slight deviations from it, in gold and silver, at current-densities of the order of 10^6 *amps* cm^2) we have to discard the idea. The lesson is, that the electrons must be supposed to be normally in motion at speeds enormously greater than the speed imparted by the field during a free fall. Let u stand for the natural average speed of the electrons; we have $T = l/u$, and

$$\sigma = \tfrac{1}{2} ne^2 l/mu, \tag{3}$$

provided always that $u \gg eET/m$.

This condition is abundantly fulfilled if we make the obvious and appealing assumption that the electrons are moving with the same average kinetic energy as atoms of a gas at the same temperature; in fact, if the free path l is no longer than the average distance between atom-centres, the deviations from Ohm's law should not appear even under such extreme circumstances as those of Bridgman's experiments. Making therefore this assumption, which in symbols is $\tfrac{1}{2}mu^2 = \tfrac{3}{2}kT$, we find

$$\sigma = \tfrac{1}{2}\sqrt{\frac{e^2}{3km}}\,\frac{nl}{\sqrt{T}}. \tag{4}$$

Not much attention should be paid to the numerical factor, which would be slightly different if we should assume Maxwell's law of distribution for the velocities of the electrons; the essential factor is the last one, $nl/\sqrt{T}$. Examining (4) in the light of the fact that the conductivity of most metals decreases distinctly more rapidly than $1/\sqrt{T}$ = in fact, as rapidly as $1/T$ or still more so — as the temperature increases, we see that the product nl will have to be supposed to vary with temperature. It seems natural to suppose that l depends altogether on the distance between adjacent atoms, which varies comparatively little with temperature, and anyway varies in the wrong direction for the purpose of the theory; so that the burden of accounting for the proportionality of σ to the first or a higher power of $1/T$ must be laid upon n.

Now it has occurred to a number of people that the free electrons are dissociated from the atoms, and the number of free electrons

is given by the degree of dissociation, which in turn should vary with the temperature in a manner prescribed altogether by the amount of work necessary to remove an electron from an atom into the (presumed) interspace where it plays about freely. But we should certainly expect that this work would be positive, as it is for the extraction of electrons from free atoms; in which case the degree of dissociation and the number of free electrons should increase with temperature. The theory is therefore adapted to explain a resistance which decreases steadily with increasing temperature, as do the resistances of some non-metallic elements; it is adapted to explain a resistance which at first diminishes and then, as the temperature increases further, goes through a minimum and rises, for the decrease in the factor $l \sqrt{T}$ finally predominates over the increase in the factor n; it is not adapted to explain a resistance increasing with temperature over the whole range, as do those of the metals. One might assume that the work of extracting an electron from an atom inside the metal is negative. This is essentially the alternative embraced by Waterman, who postulates that the work in question is a function of temperature, of the form $W = W_0 - cT$, $c > 0$. For metals W is to be chosen negative or zero, so that W shall be negative throughout; for non-metallic elements W is to be given some positive value, so that W shall change in sign at some point in the temperature-range. This unusual theory must be judged by its effectiveness; that it should reduce conduction in all elements, metallic and non-metallic alike, to a phenomenon of a single type is a feature appealing strongly in its favor; but Noyes' curves of resistance versus temperature for graphite did not agree with its demands in a satisfactory manner.

The assumption underlying (1) has however involved us in a collateral difficulty. If we believe that the n free electrons per cc. of the metal have an average energy $\frac{3}{2}kT$ and a total kinetic energy $\frac{3}{2}nkT$, we are certainly forced to admit that when the unit cube of metal is heated through 1° the electrons must take their share $\frac{3}{2}nk$ of the heat imparted to it; but the specific heat of most metals is such that it seems that the atoms must take it all and leave none over for the electrons. If we evade this difficulty by assuming n to be quite small compared with the number of atoms per cc., a few per cent. of it or less, we lose certain numerical agreements which will be mentioned later, and we have also to make l quite large, amounting to several times the distance between adjacent atoms;

yet all the tendency of modern atomic theory is to make it seem likely that the atoms fill almost the whole space within the metal.

Another way to avoid the difficulty with the specific heats consists in assuming that the high natural speed with which the electrons fly about is altogether independent of temperature; the burden of making σ as expressed in (3) vary in the proper manner with temperature is then laid upon l, which, Wien suggested, should be supposed to vary inversely as the amplitude of vibration of the atoms—that is, a free electron collides with an atom only if and when it is in vibration, and the chance of a collision increases with the amplitude of the vibration. The variation of resistance with pressure may then be explained, so far as the usual sign goes, by saying that when an ordinary metal is compressed the amplitude of oscillation of its atoms diminishes, though the temperature remain the same; the frequency of oscillation must then vary inversely as the amplitude, to keep the average energy of oscillation constant; there is some reason for expecting this to happen. Bridgman's theory somewhat resembles this one, except that the electrons are supposed to glide through the atoms and collide with the gaps; gaps between atoms are comparatively unusual, and occur chiefly when two atoms are vibrating with great amplitudes in opposite senses, so that the variation of conductivity with pressure again has the proper sign. But to explain the behavior of the three metals of which the resistance increases with pressure and with tension, Bridgman went back to the idea that in these the electrons glide through the interspaces.

As I have given only the phenomena of conduction which the electron-theory explains with difficulty, I must in justice mention the ones on which its reputation chiefly depends. In the first place it is a theory of thermal conduction as well as electrical conduction; the electrons in the hotter part of a metal maintained at an uneven temperature are assumed to have a greater average energy than the electrons in the cooler part, so that they diffuse down the temperature-gradient and realize a convection-current of heat. The theory leads to as definite a numerical value of the one conductivity as of the other, and the ratio of electrical to thermal conductivity is predicted as

$$\frac{\lambda}{\sigma} = 2\left(\frac{k}{e}\right)^2 T, \tag{5}$$

a universal constant for all metals, multiplied into the absolute temperature, and devoid of the quantities n and l which have caused us so much trouble. This is one of the predictions which is nearly enough true to be impressive; the ratio $\lambda \div \sigma T$ does indeed vary sur-

prisingly little over a wide range of metals at room-temperature and over a fairly wide range of temperatures for each of many metals. It is usually somewhat larger than the predicted value (5); but this can be conveniently explained by saying that there must be an additional mechanism for transmitting heat, something in the nature of the elasticity of the substance, which superposes its conducting-power upon the conducting-power of the electrons, and so inflates the numerator of the ratio in (5). The reason for supposing such an extra mechanism is primarily that there must be some such mechanism to perform the thermal conduction in substances which are electrical insulators. No element conducts heat as badly as sulphur and boron conduct electricity; and if we imagine a special elastic mechanism for conducting heat in boron and sulphur, we can hardly deny it to copper and silver. Bridgman found that for six metals out of eleven tested, the thermal conductivity decreased when high pressure was applied, although the electrical conductivity increased. We must hope to find an explanation for this anomaly in the behaviour of the elastic mechanism; likewise an explanation for the deviations from (5) which occur at high and at low temperatures. In theories such as the one mentioned over Wien's name in the last paragraph, in which the average *vis viva* of the electrons is supposed not to vary from a hotter place in a metal to a cooler place, we have to lay the entire burden of thermal conduction upon the elastic mechanism. This makes it difficult to explain the universal relation (5).

Another striking feature of the theory is that Lorentz succeeded in deducing the Rayleigh-Jeans radiation-law from it. He obtained from it an expression for E, the radiant emissivity of a thin stratum of metal, as a function of temperature T of the metal and wavelength λ of the radiation; another for A, the absorbing-power of the metal, likewise a function of T and λ; divided the first by the second, and obtained a definite quotient. By Kirchhoff's thermodynamic laws, E/A is equal to E_λ, the radiant emissivity of a perfectly black body. The expression deduced by Rayleigh and by Jeans for E_λ and the expression deduced by Lorentz for E/A are identical. Lorentz assumed that the collisions of the electrons with the atoms (or whatever it is they collide with) are very short in duration compared with the intervals of free unaccelerated flight from one collision to the next, and that the speeds of the electrons are distributed according to Maxwell's law about the mean value corresponding to the mean energy $3kT/2$. He also made certain assumptions which restrict the validity of his expression for E/A to radiations of great wavelength; the Rayleigh-Jeans expression for E_λ is restricted in exactly

the same way. At least as much, it seems, should be demanded from any theory of conduction offered in competition with the "classical" one.

The conception of free electrons in metals also gives a beautiful qualitative explanation of the thermoelectric effects, although unfortunately it does not do very well as a quantitative theory. If in two metals at a certain temperature the densities of free electrons are different — n_1 free electrons per cc. in one and n_2 in the other — and these two metals are brought into contact with one another, electrons will flow from the one where the density is greater into the one where it is less; and this flow will continue until arrested by a counter-electromotive-force V, of which the equilibrium-value can be shown, in any one of a variety of ways, to be

$$V = \frac{kT}{e} ln(n_1 \ n_2)$$

Such an electromotive force would account for the Peltier effect; and conversely, if the theory were correct, measurements of the Peltier effect between two metals at a given temperature and pressure would give the ratio between the densities of free electrons in the two metals under the specified conditions. Such data, combined with data on conductivity interpreted by such an equation as (4), should give information about the free paths l_1 and l_2 in the metals. The Thomson effect is more difficult to deal with, as thermal equilibrium does not prevail; however it can be seen that there will be a counter E.M.F. in an unevenly-heated metal. Measurements on the Peltier and Thomson coefficients for many metals, over wide ranges of temperature and pressure, would be very valuable; but they are so extremely hard to make even under the best of conditions, that the outlook for obtaining a really extensive set is unpromising. Possibly there is a better chance with the indirect method (determining the first and second derivatives of the curve of thermal electromotive force versus temperature). Such data of the Thomson effect as exist are not helpful to the simple theory.

Another phenomenon which lends itself very readily to explanation by the theory, and so contributes a certain amount of support to it, is the thermionic effect — the spontaneous outflow of electrons through the surfaces of hot metals. (But carbon likewise exhibits it very efficiently, and we must beware of formulating any theory of it which reposes on specific properties of metals not shared by carbon!) To interpret the thermionic effect only one new feature need be added to the theory, and this a feature which in fact was all the time latent

in it — the idea that there is a certain fixed potential-difference between the interior of a metal and the region outside of it, resulting in a potential-drop localized in a thin stratum at the surface, which an electron within the metal must surmount in order to escape from the metal into a contiguous vacuum. Such a potential-drop would for instance result from a "double layer" along the surface of the metal, a sheet of positive charges within and a sheet of negative charges opposite, parallel, and close to the positive sheet on the outside. It has been pointed out that, since probably half of the orbital electrons belonging to the atoms at the frontier of a metal lie outside the plane containing the nuclei of these atoms, they with the nuclei constitute a sort of double-layer; it has also been suggested that after a certain number of electrons issue from the metal, they are held as an electron-atmosphere above it by the forces due to the distribution of residual positive charge within the metal (Kelvin's electrical-image conception), and the electron-atmosphere with the positive surface-charge together form a double-layer. However we may conceive this double-layer, it is obvious that if we postulate free electrons within the metal, we must also postulate a barrier in the shape of an opposing potential-drop between the metal and the exterior world to keep the electrons from wandering away.

Designate this potential-drop by b, so that eb is the energy which an electron must give up in traversing it from inside to outside. Assume further (disregarding the old specific-heat difficulty) that the velocities of the electrons inside the metal are distributed isotropically in direction, and according to Maxwell's distribution-law in speed, with the mean kinetic energy $\frac{3}{2}kT$ appropriate to the temperature T of the metal. Imagine the metal surface to occupy the plane $x=0$, metal to the left and vacuum to the right. Consider the electrons which come from within the metal and strike unit area of the boundary in unit time; those of them which have velocities of which the x-component lies between u and $u+du$ are in number equal to

$$dI = \frac{nu}{\sqrt{2\pi kT/m}} e^{-\frac{mu^2}{2kT}} du, \tag{6}$$

n meaning as heretofore the number of electrons per unit volume of metal. The total number which strike unit area of the boundary from within is equal to the integral of this expression from $u=0$ to $u=\infty$, which is

$$I_0 = n\sqrt{kT/2\pi m}. \tag{7}$$

Those which escape are those for which $\frac{1}{2}mu^2$ exceeds eb; we obtain the number of them by integrating (6) from $u=\sqrt{2eb/m}$ to $u=\infty$, and find

$$I_e = n\sqrt{kT/2\pi m}\, e^{-eb/kT}. \tag{8}$$

This, supposing n and b to be independent of temperature, is Richardson's well-known formula for the saturation-current from a hot body as function of temperature. All of the multitudinous observations agree with it; but this does not mean so much as might be thought, for the experts inform us that all the data, no matter how accurately taken, would agree quite as well with a formula in which T, or T^2, or even T^0, stood in the place of the factor $T^{1/2}$ by which the exponential is multiplied. Incidentally this would permit us to make n vary as some small power of temperature, such as the inverse square root, if we chose to make the resistance-temperature relation in (4) agree with experiment at such a price. Or if we assume n independent of temperature, we can calculate it from measurements on thermionic saturation-currents. The measurements usually give for n values of the order of magnitude of the number of atoms per unit volume.

What is more definitely significant is, that the velocities of the emerging electrons are actually distributed in a manner compatible with the assumptions made. Let us enquire how many of the electrons issuing from unit area of the metal have velocities of which the x-component lies between u and $u+du$. These are the very same electrons which struck the surface from within, having velocities of which the x-component lay between u' and $u'+du'$; u' and du' being related to u and $u+du$ by the equations:

$$\tfrac{1}{2}mu^2+eb=\tfrac{1}{2}m(u')^2, \quad u'du'=udu. \tag{9}$$

The number of these electrons is by (6)

$$dI' = \frac{nu'}{\sqrt{2\pi kT/m}}\, e^{-\frac{mu'^2}{2kT}}\, du', \tag{10}$$

which by virtue of the relations (9) reduces to

$$e^{-\frac{eb}{kT}}\, \frac{nu}{\sqrt{2\pi kT/m}}\, e^{-\frac{mu^2}{2kT}}\, du, \tag{11}$$

which is identical with (6) except for a constant factor; which means in turn that the distribution-function of the emerging electrons is identical with the distribution-function of the internal electrons,

being in fact the Maxwell distribution-function with the same mean kinetic energy $\frac{3}{2}kT$. The argument as given proves the point only for the distribution in the velocity-component u; but the distribution-functions in v and w, the velocity-components parallel to the boundary of the metal, are unaffected by the double-layer, since v and w for any particular electron are unaffected by the passage through it; and since it is the essential feature of the Maxwell distribution-law that the distributions in v and w are identical for each and every value of u, the conclusion follows as stated. Nevertheless it does sound paradoxical.

This conclusion has been verified repeatedly by experiment. Richardson began by simulating the simple mathematical conditions of infinite plane electrodes as closely as practicable; he inserted a small flat incandescent surface in an aperture in the middle of a large flat cold plate, charged the two to the same potential, and placed opposite and parallel to them a large flat collecting-electrode. Charging this latter to various potentials V inferior to the potential of the emitting surface, he plotted the electron-current which it received as function of V; this is the distribution-function of the speed u of equation (6) and the following equations translated into terms of the corresponding kinetic energy $\frac{1}{2}mu^2$ as independent variable. To ascertain the distribution-functions in v and w he isolated a small area of the collecting-electrode, moved it to and fro in a plane parallel to the plane of the emitting surface, and measured the current into it in its various positions. Many measurements have since been made upon the currents into cylindrical collectors from hot wires stretched along the axes of the cylinders; it is somewhat more difficult to write out the formula for the expected relation between current and retarding-potential, but the experimental conditions are much more under the experimenter's control. All these investigations have confirmed the theorem, except a single discordant one which was later explained away; the strongest verification is furnished by the experiments of Germer, whose precautions of preparation and accuracy of measurements far surpassed everything that had gone before.

The evidence thus is quite favorable to the idea of an electron-gas within the metal with its electrons moving with velocities as prescribed by Maxwell's distribution-law, and kept from diffusing away by a double-layer covering the surface. Other evidence for the existence of a double-layer is furnished by the photoelectric effect and by the existence of contact-potential-differences. When

light of frequency ν falls upon a metal, electrons emerge from it with velocities which are distributed in a manner quite distinct from Maxwell's distribution and have nothing to do with the temperature of the metal. The kinetic energies of some of the electrons attain a certain upper limit W_m, but none surpasses it; W_m is a linear function of ν given by the equation

$$W_m = h\nu - P, \tag{12}$$

h being Planck's constant, P a positive constant characteristic of the metal. This is an exceedingly strong intimation that each of the emerging electrons, while still inside the metal, suddenly absorbed a quantum of energy $h\nu$ from the light and departed with it, giving up a fixed quantity P in passing through the surface. (Those which issue with energies clearly less than W_m can be supposed to have started distinctly beneath the surface and to have lost additional energy in struggling through the metal to it). Translating P into potential-drop, we see that it represents the potential-difference or the "strength" of the surface double layer. It may be determined by measuring W_m for light of various frequencies, plotting it against frequency, and extrapolating the resulting straight line to its intersection with the axis of frequencies. Or it may, in principle, be determined by plotting the photoelectric current as a function of frequency, and extrapolating the curve to its intersection with the axis of frequencies, where no electrons escape and the photosensitiveness ceases; but curves are not so easy to extrapolate as straight lines, and there are some anomalous results which are still unexplained.

It would seem an easy matter to measure the strength of the double-layer by both photoelectric and thermionic methods upon a single substance. But it is rather difficult; for one reason, the substances for which the photoelectric currents are easy to produce and measure are precisely the metals upon which good thermionic measurements are next to impossible, and vice versa. The best photoelectric measurements have been made upon the alkali metals, which are very sensitive to visible light; but they cannot be formed into wires, and volatilize furiously when heated enough to produce an important thermionic effect, filling the evacuated tube with dense vapors which ruin the accuracy of the measurements. The best thermionic measurements have been made upon platinum and tungsten, which are not sensitive at all to visible light, and begin to be sensitive far out in the ultraviolet where experiments with radiation are difficult. Furthermore there is the capital difficulty that the photoelectric measurements must be confined to temperatures where the thermionic current

is imperceptible; if one were to irradiate an incandescent tungsten filament the extra current of photoelectrons would be too small to notice. If we assume outright that P does not vary greatly from room-temperature up to the temperatures of incandescence, and therefore compare photoelectric data upon cool metals with thermionic data upon the same metals when hot, we find that there is a fairly good agreement. Values of the thermionic constant b between 4 and 5 volts correspond to photoelectric sensitiveness commencing between 3,400 and 2,500 Angstrom units, and this correctly describes the behavior of several of the heavy high-melting-point metals; photoelectric sensitiveness extending well up into the visible spectrum, such as the alkali metals display, corresponds to values of $P\ e$ of the order of 2 volts and lower, and such values are indicated by the thermionic experiments made upon sodium and potassium by Richardson under the inevitably bad conditions.

Contact-potential-difference, one of the longest known of all electrical phenomena Volta discovered it agrees admirably with this interpretation of the photoelectric constant P. Imagine that we have pieces of two metals, potassium and silver for example, which are drawn out and welded together at one end, and at their other ends are spread out into plates and face one another across a vacuous space. We know that the opposing faces behave as if they were at essentially different potentials, the potential-difference V between them being characteristic of the two metals and independent of the size or separation of the opposing faces. Yet this potential-difference V is not equal to, is indeed usually much greater than the potential-difference between the interiors of the metal across the welded joint, which is deduced from the Peltier effect. The only way to resolve the contradiction is to assume that it is the region just outside the potassium which differs by V from the region just outside the silver; the metals themselves are at nearly the same potential, but there is a double-layer at the surface of each which establishes a fixed potential-drop between it and the vacuum. Representing by $P_1\ e$ and by $P_2\ e$ the voltage-drops at these two double-layers, by M the potential-difference between the interiors of the metals as inferred from the Peltier effect, by V the potential-difference between the regions just outside the metals which we identify with the contact potential difference, we find

$$P_2\ e - P_1\ e = V + M \tag{13}$$

in which M is so small compared with the other terms that henceforth we will leave it out.

Now imagine that light of a high frequency ν_0 falls upon the potassium; it elicits electrons of which the maximum energy at emergence is $h\nu_0 - P_1$; these highest-speed electrons arrive at the silver plate with energy $(h\nu_0 - P_1 - e - V)$, having had to overcome the additional potential-drop V in passing from the region just outside the potassium to the region just outside the silver. (The reader can make the changes in language required if V happens to be of the sign corresponding to a potential-rise). From (13) we see that this energy of arrival is equal to $(h\nu_0 - P_2 - e)$—an expression from which P_1, the only quantity characterizing the irradiated metal, has fallen out! Therefore the electrons arrive at the silver plate with the same maximum speed, whether the irradiated metal be potassium, sodium, silver, or any other metal! (unless we hit upon a metal for which $h\nu_0 < P_1 - e$, in which case we shall never get any at all).

This experiment is usually performed by putting a battery between the silver and the irradiated metal, and adjusting its E.M.F. until the fastest electrons are just turned back before reaching the silver; this is known as "determining the stopping potential." If our interpretation of contact-potential-difference is correct, the stopping-potential must be independent of the irradiated metal, and depend only on the material of which the collecting-electrode is made; further, the difference between the stopping-potentials observed with two different metals as collecting-electrodes should be equal to their contact potential difference. These predictions have been verified in several sets of experiments, notably by Richardson and Compton. Millikan developed the interesting theoretical consequences which they suggest. There should be similar relations involving thermionic currents; observations confirming them have been made, but not so extensively published; they are more difficult to make with accuracy because the thermionic electrons have no definite recognizable maximum velocity.

We seem to have marshalled a formidable amount of evidence in favor of the electron-theory of conduction with the associated idea of the surface double-layer. Yet it would be misleading not to point out that an equation quite as satisfactory as (8) in representing the thermionic current as function of temperature can be deduced by reasoning in an entirely different fashion from entirely different postulates. This, the thermodynamical method of speculating about the thermionic effect, was originated by H. A. Wilson; it consists essentially in assuming a thoroughgoing analogy between the outflow of electrons from a hot metal and the evaporation of molecules from a solid or a liquid. We know that if an evacuated chamber is partly

filled with liquid water or solid $CaCO_3$, the remaining space inside the chamber is quickly pervaded with H_2O or CO_2 molecules composing a gas, its pressure and density being determined absolutely by the temperature T. We infer that if an evacuated chamber, with its walls made of some insulating substance, contains a piece of metal and is heated to a high temperature, the whole evacuated space will be pervaded with electrons composing a gas, its pressure p and density n being determined absolutely by the temperature of the system, T. We must assume that the electron-gas outside the metal conforms to the ideal-gas law

$$p = nkT \tag{14}$$

and we shall also presently assume that its specific heats have the values characteristic of monatomic ideal gases,

$$C_v = \frac{3}{2}Nk, \qquad C_p = \frac{5}{2}Nk. \tag{15}$$

I use n to represent the number of electrons per unit volume of the gas, as the number within the metal no longer enters in any way into the reasoning; N to represent the number in a gramme-molecule (Avogadro's constant). These are the only assumptions which involve a kinetic theory in any way.

Imagine now a wire of which one end projects into an evacuated chamber of the sort described, maintained at T, and the other into another such chamber maintained at $T+dT$. We consider a process which consists of increasing the volume of the first chamber by just enough to require N additional electrons to come out of the wire to fill the additional space, and simultaneously decreasing the volume of the second chamber by just enough to crowd N electrons into the wire; so that in effect N electrons are transferred from the one chamber to the other through a wire of which the two ends are at temperatures $T+dT$ and T. This process will be carried on reversibly. Designate by L the heat which must be imparted to the metal at T, to remove one electron from it under the circumstances of the experiment; by sdT the heat which is absorbed when one electron is transferred through the metal from a point where the temperature is T to a point where the temperature is $T+dT$. s is the coefficient of the Thomson effect, referred to a single electron instead of a coulomb. L contains a term kT, which corresponds to the mechanical work done in forcing back the walls of the chamber to make place for the evaporated electron-gas. Subtracting it we

obtain $(L - kT)$, to be called $e\phi$, as the actual energy expended in putting the electron across the boundary of the metal.*

In the process which I have just described, the *input of heat* consists of the following terms: NL which goes to extract the N electrons from the metal in the first chamber, $-N\left(L+\frac{dL}{dT}dT\right)$ which is liberated when N electrons condense into the metal in the second chamber, and $-NsdT$ which is absorbed by the electrons in travelling through the wire. The *output of work* is NkT during the expansion of the first chamber, $-NkT-NkdT$ during the contraction of the second chamber. The *input of entropy* is NL/T during the evaporation in the first chamber, $-N\left(L/T+\frac{d(L/T)}{dT}\right)dT$ during the condensation in the second chamber, and $(-Ns/T)dT$ during the flow of electrons through the wire.

We now complete the cycle by changing the pressure and temperature of the gramme-molecule of electron-gas in the first chamber from p, T to $p+dp$, $T+dT$, after which it becomes equivalent with the gramme-molecule in the second chamber at the beginning of the process. Calculated in the usual way isothermal contraction at T from p to $p+dp$, isobaric expansion at $p+dp$ from T to $T+dT$ we find: *input of heat*, $\frac{5}{2}NkdT-NkT[d(\ln p)/dT]dT$; *output of work*, $NkdT-NkT[d(\ln p)/dT]dT$; *input of entropy*, $\frac{5}{2}(Nk/T)dT-Nk[d(\ln p)/dT]dT$.

The two processes together constitute a complete reversible cycle. We therefore equate the sum of the inputs of entropy to zero, and obtain:

$$\frac{e\phi}{T^2}\frac{d\phi}{dT}-s+\frac{5}{2}k-kT\frac{d(\ln p)}{dT}=0. \tag{16}$$

and equate the difference of the inputs of heat and the outputs of work to zero, which gives:

$$e\frac{d\phi}{dT}-s+\frac{3}{2}k=0 \tag{17}$$

* This definition suggests a thermal method of measuring L, which has several times been put into practice. The experiments are difficult and the data must be corrected for many influences, but the best results indicate that $(L-kT)$ is approximately equal to $e\phi$ of (8). The data of Dorisson and Germer indicate a slight difference, which may be an important test of suggested theories of conduction.

and combine the equations into

$$\frac{e\Phi}{T} + k = kT\frac{d(\ln p)}{dT} \tag{18}$$

which integrated, yields

$$p = AT e^{\int \frac{e\Phi}{kT^2} dT}. \tag{19}$$

We still have to make the bridge between this formula, which relates to the pressure of the electron-gas in equilibrium with the metal, and the quantity actually observed, which is the saturation-current out of the metal surface in an accelerating field. In the equilibrium-state, the number of electrons which issue from the metal is equal to the number which, coming from the external electron-gas, strike its boundary and do not rebound. This is indisputable; to make it useful we have to make two new assumptions: one, that the number of electrons which issue from the metal is the same in an accelerating field as in the equilibrium-state; the other, that no electrons rebound from the surface. The first assumption had to be made in the preceding deduction — that is, we had to assume tacitly that the uncompensated outflow of electrons through the surface of the metal did not appreciably distort the Maxwell distribution within; the second is a drawback peculiar to the thermodynamic method. Accepting these two assumptions along with all their predecessors, we finally reach the expression for the number of electrons emitted per unit area per unit time from the surface of the hot metal:

$$I = CT^{\frac{1}{2}} e^{\int \frac{e\Phi}{kT^2} dT}. \tag{20}$$

This is the equation for the thermionic saturation-current attained by the thermodynamical reasoning.

Let us finally try some hypotheses about the variation of Φ with temperature: for a first one, the hypothesis $\Phi = \text{constant}$. The general equation becomes

$$I = CT^{\frac{1}{2}} e^{-\frac{e\Phi_0}{kT}}, \tag{21}$$

which is perfectly identical with (8) which was deduced from the electron-theory with the additional assumption of a double-layer independent of temperature. We cannot however freely make an assumption like this, for our equation (17) shows that an assumption about $d\Phi/dT$ implies, and conversely is implied by, an assumption

about the value of the Thomson coefficient s. In making ϕ independent of temperature we in effect assumed that the Thomson coefficient has the value $s = \frac{3}{2}k$ (per electron), which happens to be precisely the value demanded (and vainly demanded) by the electron-theory of conduction. If on the other hand we choose to accept from the experiments the fact that s is extremely small compared to $\frac{3}{2}k$, the equation (16) compels us to set

$$\phi = \frac{3}{2}kT \quad c + \phi_0. \tag{22}$$

Inserting this into (19) we obtain

$$I = CT^2 e^{-\frac{e\phi_0}{kT}}, \tag{23}$$

which is commonly known as the T^2-law, and is at the moment the favorite way of expressing the variation of thermionic current with temperature. As I said earlier, experiment is thus far powerless to distinguish between (8), (20) and (22).

This brief and superficial sketch of the thermodynamic argument is meant partly to familiarize the reader with the T^2 formula, and partly to show that the observations upon the dependence of thermionic current on temperature do not necessarily sustain the particular type of theory which has figured most in these pages, as against its rivals actual or conceivable. Of course it would be unjustifiable to say that any argument of the thermodynamical type is *ipso facto* stronger than any argument based on a physical model. It may be true that the laws of thermodynamics are valid everywhere without exception; but it is certainly true that in any particular case it is extremely difficult to feel sure just how they should be applied to arrive at absolutely binding conclusions. In this case, for instance, we have assumed as both possible and reversible a process which no one has ever carried through, and no one, in all likelihood, ever will; and in the course of analyzing the transfers of energy between the system and the external world in this imagined process, we have classified some as transfers of heat and some as transfers of mechanical work, and possibly ignored yet others, so that the analysis requires careful thought and has in fact been made in different ways by different authorities. There is for example the problem of the allowance to be made for work done in transferring the electrons from place to place against electromotive forces, which might or might not be *nil* when summed around the complete cycle; H. A.

Wilson has recently made a specific assumption regarding these. For still further subtleties Bridgman's theoretical articles may be consulted. I must however add that an extension of the thermodynamical argument, with the assistance of Nernst's "third law of thermodynamics," leads to the conclusion that the constant C of equation (23) should have for all elements, if not indeed for all substances, the same universal value, calculable in terms of certain universal constants. There is some evidence that this may be true for emission from pure elements. Were it so, the result would be of fundamental importance; but another article almost as long as this one would be required to explain it properly.

The general tone and character of this article will probably leave the final impression that the electrical behaviour of solids is an utterly confused and chaotic department of physics, a hopeless entanglement of incongruous rules diversified by numberless exceptions. I fear that this impression — except perhaps for the hopelessness of the situation — is substantially the correct one. In fact this presentation has put the state of affairs in rather too favorable a light, for I have passed over a number of the perplexities. I have scarcely mentioned the thermoelectric effects, or spoken of the complexities of the photoelectric effect, or of the emission of electrons from metals which are bombarded by other electrons or by ionized atoms; and I have not mentioned at all the galvanomagnetic and thermomagnetic effects, the most baffling and bewildering of all. In fact it seems only too probable that if one should succeed in erecting a theory by which all the phenomena I have described could be brought into one coherent system, some galvanomagnetic effect would be lying in wait for it to bring it to the dust. Clairaut is said to have been saddened by feeling that Newton had discovered all the laws of celestial mechanics, leaving nothing for men born after him to do except to improve the methods of calculation. Ambitious students of physics who, through too exclusive a study of the radiations from atoms, may have come to feel in the same way about Bohr, should find consolation in contemplating the present status of the Theory of Conduction in Solids.

LITERATURE

The chief recent compilation of data upon conduction in solids is Koenigsberger's article in Graetz' *Handbuch der Elektrizität*. K. Baedeker wrote an excellent short account of the data and the theories, entitled *Die elektrischen Erscheinungen in metallischen Leitern*, which although published in 1911 is not yet superseded. Bidwell's paper on germanium is in Phys. Rev. (2) *19*, pp. 447–455 (1922). Noyes' article on carbon in Phys. Rev. (2) *24*, pp. 490–499 (1924). The investigations on supraconductivity are reported chiefly in the *Leyden Communications*, Commelin

has given a comprehensive account of them in Phys. ZS. *21* (1920), with a bibliography of all of the work; two or three subsequent communications are reviewed in Science Abstracts. Bridgman's work on the effect of pressure and of tension on the electrical and thermal conductivities of the elements is printed chiefly in the *Proceedings of the American Academy of Arts and Sciences* from 1917 onward, with occasional announcements in the Physical Review, where also his theoretical papers are published (Phys. Rev. *14*, pp. 306–347 (1919); *17*, pp. 161–195 (1921) and *19*, pp. 114–134 (1922). For the effect of melting, consult Bridgman's papers, and one by von Hauer, in Ann. d. Phys. *51*, pp. 189–219 (1916).

The "classical" theory of conduction is presented in Lorentz' book *The Theory of Electrons*, which bears his signature as of 1915. Bohr wrote a dissertation upon it which is highly praised by those who have succeeded in reading it in the Danish. Wien's and Planck's modifications of it are published in the *Sitzungsberichte* of the Berlin Academy for 1912 and 1913. In the Philosophical Magazine of 1915 there are a number of articles on the theory by G. H. Livens, like Baedeker a victim of the war. The conception of quantity of free electrons determined by dissociation of atoms is presented by Koenigsberger in Ann. d. Phys. *32*, pp. 170–230 (1910) and Waterman's extension of it is in Phys. Rev. *22*, pp. 259–270 (1923). Some chapters in J. J. Thomson's *Corpuscular Theory of Matter* deal with the theories; in an article in Phil. Mag. *29* (1915) he offers a theory involving an attempt on supra-conductivity, which the others do not touch.

The field of thermionics is thoroughly covered in Richardson's *Emission of Electricity from Hot Bodies* (2d edition, 1921). Subsequent theoretical papers by Richardson are in Proc. Roy. Soc. *A105*, pp. 387–405 (1924) and Proc. Phys. Soc., London, *36*, pp. 383–399 (1924), and one by H. A. Wilson on what I have called the "thermodynamical argument" in Phys. Rev. (2) *24*, pp. 38–48 (1924). For various interpretations of the surface double-layers see Debye, Ann. d. Phys. *33*, pp. 440–489 (1910); Schottky, ZS. f. Phys. *14*, pp. 63–106 (1923); and Frenkel, Phil. Mag. 33, pp.297–322 (1917). Germer's investigation of the distribution-in-energy of thermionic electrons is briefly reported in Science, *42*, 392 (1923), and a fuller account is to be published; Davisson and Germer's determination of L in Phys. Rev. *20*, pp. 300–330 (1922). For the photoelectric measurements establishing equation (12), consult Millikan, Phys. Rev. *7*, pp. 355–388 (1916); for the relation between values of P and contact-potential-difference consult Page, Am. Journ. Sci. *36*, pp. 501–508 (1913) and Millikan, Phys. Rev. *7*, pp. 18–32 (1916). Values of the thermionic constant b are tabulated in Richardson's book and in Dushman's article, Phys. Rev. (2) *21*, pp. 623–636 (1923). Values of the photoelectric constant P are tabulated by Kirchner, Phys. ZS. *25*, pp. 303–306 (1921) and by Hauer (Journ. Opt. Soc. *9*, pp. 251–257 (1924). For the arguments that the constant C of equation (23) is a universal constant, consult the references given by Dushman (*l. c. supra*) and Richardson, Phys. Rev. (2) *23*, pp. 153–155 (1924); for the data, Dushman in Phys. Rev. (2) *23*, p. 156 (1924).

Theorems Regarding the Driving-Point Impedance of Two-Mesh Circuits*

By RONALD M. FOSTER

SYNOPSIS.—The necessary and sufficient conditions that a driving point impedance be realizable by means of a two-mesh circuit consisting of resistances, capacities, and inductances are stated in terms of the four roots and four poles—including the poles at zero and infinity—of the impedance. The roots and the poles are the time coefficients for the free oscillations of the circuit with the driving branch closed and opened, respectively. For assigned values of the roots, the poles are restricted to a certain domain, which is illustrated by figures for several typical cases; the case of real poles which are not continuously transformable into complex poles is of special interest. All driving point impedances satisfying the general conditions can be realized by any one of eleven networks, each consisting of two resistances, two capacities, and two self-inductances with mutual inductance between them; these are the only networks without superfluous elements by which the entire range of possible impedances can be realized; the three remaining networks of this type give special cases only. For each of these eleven networks, formulas are given for the calculation of the values of the elements from the assigned values of the roots and poles.

I. STATEMENT OF RESULTS

THE object of this paper is, first, to determine the necessary and sufficient conditions that a driving-point impedance† be realizable by means of a two-mesh circuit consisting of resistances, capacities, and inductances, and second, to determine the networks‡ realizing any specified driving-point impedance satisfying these conditions.

These necessary and sufficient conditions are stated in the form of the following theorem:

Theorem I. Any driving-point impedance S of a two-mesh circuit consisting of resistances, capacities, and inductances is a function of the time coefficient $\lambda = ip$ of the form

$$S = H\,\frac{(\lambda-\alpha_1')(\lambda-\alpha_2')(\lambda-\alpha_3')(\lambda-\alpha_4')}{(\lambda-\beta_1)(\lambda-\beta_2)(\lambda-\beta_3)} \qquad \text{(1a)}$$

$$= \frac{a_0\lambda^4 + a_1\lambda^3 + a_2\lambda^2 + a_3\lambda + a_4}{b_2\lambda^2 + b_1\lambda + b_0}, \qquad \text{(1b)}$$

* Presented by title at the International Mathematical Congress at Toronto, August 11–16, 1924, as "Two-mesh Electric Circuits realizing any specified Driving-point Impedance."

† The driving-point impedance of a circuit is the ratio of an impressed electromotive force at a point in a branch of the circuit to the resulting current at the same point.

‡ The networks considered in this paper consist of any arrangement of resistances, capacities, and inductances with two accessible terminals such that, if the two terminals are short-circuited, the resulting circuit has two independent meshes. Thus the impedance measured between the terminals of the network is the same as the driving-point impedance of the corresponding two-mesh circuit. Throughout the paper this distinction will be made in the use of the terms "network" and "circuit."

where $$H \geq 0,\ \alpha_1+\alpha_2 \leq 0,\ \alpha_1\alpha_2 \geq 0,\ \alpha_3+\alpha_4 \leq 0,\ \alpha_3\alpha_4 \geq 0,$$
$$\beta_2+\beta_3 \leq 0,\ \beta_2\beta_3 \geq 0, \tag{2}$$

and $$b_1^2(a_3^2-4a_4d)+b_2^2[(a_2-d)^2-4a_0a_4]+b_3^2(a_1^2-4a_0d)$$
$$-2b_1b_2[a_3(a_2-d)-2a_1a_4]-2b_1b_3[a_1a_3-2d(a_2-d)]$$
$$-2b_2b_3[a_1(a_2-d)-2a_0a_3]=0, \tag{3}$$

for all values of $d \geq 0$, *provided*

$$-a_4b_2^2+a_3b_2b_3-db_3^2 \geq 0, \tag{4}$$
$$-a_0b_3^2+(a_2-d)b_3b_1-a_4b_1^2 \geq 0, \tag{5}$$
$$-db_1^2+a_1b_1b_2-a_0b_2^2 \geq 0, \tag{6}$$

and, conversely, any impedance S *of the form* (1) *satisfying these conditions* (2) (6) *can be realized as the driving-point impedance of a two-mesh circuit consisting of resistances, capacities, and inductances.*

Theorem I thus gives the most general form of this type of impedance, showing that it is a rational function of the time coefficient,[3] completely determined, except for a constant factor, by assigning four roots and two poles, in addition to the poles at zero and infinity, subject to certain conditions. The assigned roots and poles are the time coefficients for the free oscillations of the circuit with the driving branch closed and opened, respectively. That is, the roots and poles correspond to the resonant and anti-resonant points of the impedance.

The conditions are as follows: The real part of each root and pole is negative or zero; the roots and poles occur in pairs of real or conjugate complex quantities; certain additional restrictions must be satisfied, as stated in terms of the symmetric functions of the roots and poles by formulas (3) (6).

By virtue of these restrictions, the pair of poles, for assigned values of the two pairs of roots, is limited to a certain domain of values. This domain is conveniently illustrated by plotting, in the upper half of the complex plane, the locus of one pole, the other pole being its conjugate. For real poles, a device is used to indicate pairs of points on the real axis. Figs. 3 5 show the domain of the poles, plotted in this manner, for several typical cases.

Provided the roots are not all real, this domain consists of a connected region of values, so that it is possible to pass from one pair of poles to any other pair satisfying the same conditions by a continuous transformation. In the case of four real roots, however, the domain consists, in general, of two non-connected regions, as illustrated in Fig. 5. Under these circumstances there is a region of real poles which are not continuously transformable into complex poles.

The networks realizing any specified driving-point impedance are

[3] All electrical oscillations considered in this paper are of the form $e^{\lambda t}$, where the time coefficient $\lambda = \alpha + i\beta$ may have any value, real or complex.

where $$H \geq 0,\ \alpha_1+\alpha_2 \leq 0,\ \alpha_1\alpha_2 \geq 0,\ \alpha_3+\alpha_4 \leq 0,\ \alpha_3\alpha_4 \geq 0,$$
$$\beta_2+\beta_3 \leq 0,\ \beta_2\beta_3 \geq 0, \tag{2}$$

and $$b_1^2(a_3^2-4a_4d)+b_2^2[(a_2-d)^2-4a_0a_4]+b_3^2(a_1^2-4a_0d)$$
$$-2b_1b_2[a_3(a_2-d)-2a_1a_4]-2b_1b_3[a_1a_3-2d(a_2-d)]$$
$$-2b_2b_3[a_1(a_2-d)-2a_0a_3]=0, \tag{3}$$

for all values of $d \geq 0$, *provided*

$$-a_1b_2^2+a_3b_2b_3-db_3^2 \geq 0, \tag{4}$$
$$-a_0b_3^2+(a_2-d)b_3b_1-a_1b_1^2 \geq 0, \tag{5}$$
$$-db_1^2+a_1b_1b_2-a_0b_2^2 \geq 0, \tag{6}$$

and, conversely, any impedance S *of the form* (1) *satisfying these conditions* (2) (6) *can be realized as the driving-point impedance of a two-mesh circuit consisting of resistances, capacities, and inductances.*

Theorem I thus gives the most general form of this type of impedance, showing that it is a rational function of the time coefficient,[3] completely determined, except for a constant factor, by assigning four roots and two poles, in addition to the poles at zero and infinity, subject to certain conditions. The assigned roots and poles are the time coefficients for the free oscillations of the circuit with the driving branch closed and opened, respectively. That is, the roots and poles correspond to the resonant and anti-resonant points of the impedance.

The conditions are as follows: The real part of each root and pole is negative or zero; the roots and poles occur in pairs of real or conjugate complex quantities; certain additional restrictions must be satisfied, as stated in terms of the symmetric functions of the roots and poles by formulas (3) (6).

By virtue of these restrictions, the pair of poles, for assigned values of the two pairs of roots, is limited to a certain domain of values. This domain is conveniently illustrated by plotting, in the upper half of the complex plane, the locus of one pole, the other pole being its conjugate. For real poles, a device is used to indicate pairs of points on the real axis. Figs. 3 5 show the domain of the poles, plotted in this manner, for several typical cases.

Provided the roots are not all real, this domain consists of a connected region of values, so that it is possible to pass from one pair of poles to any other pair satisfying the same conditions by a continuous transformation. In the case of four real roots, however, the domain consists, in general, of two non-connected regions, as illustrated in Fig. 5. Under these circumstances there is a region of real poles which are not continuously transformable into complex poles.

The networks realizing any specified driving-point impedance are

[3] All electrical oscillations considered in this paper are of the form $e^{\lambda t}$, where the time coefficient $\lambda = ip$ may have any value, real or complex.

230 + 234

Fold

Out

$\frac{a_0 b_2 b_3 - T}{b_1 b_2 b_3}$	
0	
$\frac{a_0 b_3^2 + T_2^2}{b_1 b_3^2}$	
$\frac{a_0 b_2^2 + T_3^2}{b_1 b_2^2}$	
$\frac{d}{b_2}$	
$\frac{T_1^2}{b_2 b_3^2}$	

determined by the arrangement and magnitudes of the elements, as given by the following theorem:

Theorem II. All driving-point impedances satisfying the necessary and sufficient conditions, as stated in Theorem I, can be realized by any one of the eleven networks shown by Fig. 1, upon assigning to the elements of each network the values given by Table I. These eleven networks are the only networks without superfluous elements by which the entire range of possible impedances can be realized.

By Theorem II, any network obtained from a two-mesh circuit

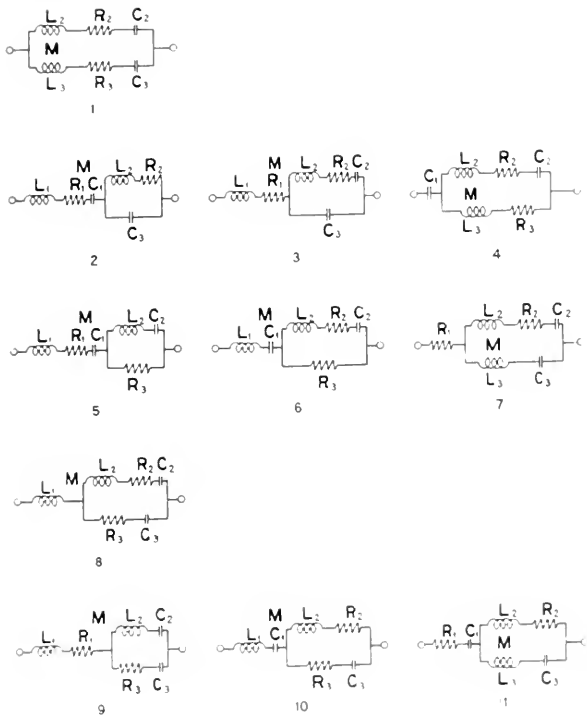


Fig. 1. Networks realizing any driving-point impedance of a two-mesh circuit consisting of resistances, capacities, self-inductances, and mutual inductances

consisting of resistances, capacities, and inductances can be replaced, in so far as the impedance between terminals is concerned, by any one of the eleven networks shown by Fig. 1, upon assigning the proper values to the elements. Each of these networks consists of two resistances, two capacities, and two self-inductances with mutual inductance between them.

Each of these eleven networks realizes impedances with arbitrarily assigned roots and with poles anywhere in the entire domain of possibilities, subject to the general conditions stated in Theorem I. Special cases of these networks realize, for arbitrarily assigned roots, only critical lines and points in the domain. All these special cases are listed in Table III, with a specification of the lines or points in the domain realizable by each, as illustrated by Figs. 4 and 5.

Certain limited regions of the domain can be realized by networks which contain no mutual inductance and which are not special cases of the networks given by Theorem II. These networks are given by the following theorem:

Theorem III. Any driving-point impedance of a two-mesh circuit consisting of resistances, capacities, and self-inductances can be realized by at least three and not more than five of the twelve networks shown by Fig. 2, upon assigning to the elements of each network the values given by Table II. These twelve networks are the only networks without mutual inductance and without superfluous elements by which any impedance can, in general, be realized.

These twelve networks, taken together, cover that portion of the domain realizable without mutual inductance. Networks with mutual inductance are needed in order to cover the entire domain. These twelve are the only networks, without superfluous elements, realizing limited regions in the domain. Each of these networks consists of two resistances, two capacities, two self-inductances, and one additional resistance, capacity, or self-inductance. The twelve networks, with their special cases, are all listed in Table III, with a specification of the regions, lines, or points realizable by each.

In addition to the specific formulas for the networks of Figs. 1 and 2, it is convenient to have general formulas for the computation of all networks meeting the given conditions, including those networks with superfluous elements as well as all special cases. The most general two-mesh circuit is shown by Fig. 6; accordingly, the most general network under consideration is that shown by Fig. 7. Formulas

consisting of resistances, capacities, and inductances can be replaced, in so far as the impedance between terminals is concerned, by any one of the eleven networks shown by Fig. 1, upon assigning the proper values to the elements. Each of these networks consists of two resistances, two capacities, and two self-inductances with mutual inductance between them.

Each of these eleven networks realizes impedances with arbitrarily assigned roots and with poles anywhere in the entire domain of possibilities, subject to the general conditions stated in Theorem I. Special cases of these networks realize, for arbitrarily assigned roots, only critical lines and points in the domain. All these special cases are listed in Table III, with a specification of the lines or points in the domain realizable by each, as illustrated by Figs. 4 and 5.

Certain limited regions of the domain can be realized by networks which contain no mutual inductance and which are not special cases of the networks given by Theorem II. These networks are given by the following theorem:

Theorem III. Any driving-point impedance of a two-mesh circuit consisting of resistances, capacities, and self-inductances can be realized by at least three and not more than five of the twelve networks shown by Fig. 2, upon assigning to the elements of each network the values given by Table II. These twelve networks are the only networks without mutual inductance and without superfluous elements by which any impedance can, in general, be realized.

These twelve networks, taken together, cover that portion of the domain realizable without mutual inductance. Networks with mutual inductance are needed in order to cover the entire domain. These twelve are the only networks, without superfluous elements, realizing limited regions in the domain. Each of these networks consists of two resistances, two capacities, two self-inductances, and one additional resistance, capacity, or self-inductance. The twelve networks, with their special cases, are all listed in Table III, with a specification of the regions, lines, or points realizable by each.

In addition to the specific formulas for the networks of Figs. 1 and 2, it is convenient to have general formulas for the computation of all networks meeting the given conditions, including those networks with superfluous elements as well as all special cases. The most general two-mesh circuit is shown by Fig. 6; accordingly, the most general network under consideration is that shown by Fig. 7. Formulas

Table I

Networks	1	2	3	4	5	6	7	8	9	10	11
M	$\frac{a_0T_1^2-dT_2^2-a_4T_3^2-a_3T_2T_3}{b_1T_1^2}$	$\frac{T_1T_3}{b_2^2b_3}$	$\frac{(a_3b_2-db_3)T_3}{b_2^2T_1}$	$\frac{(cb_3-a_4b_1)T_1+(a_3b_3-a_4b_2)T_2}{b_3^2T_1}$	$\frac{T_1T_2}{b_2b_3^2}$	$\frac{(a_3b_3-a_4b_2)T_2}{b_3^2T_1}$	$\frac{(a_1b_2-db_1)T_1+(a_3b_2-db_3)T_3}{b_2^2T_1}$	$\frac{-b_1U_1^2\pm(a_1b_3-cb_2)U_1}{2T_1^2}$	$\frac{(db_3-a_3b_2)(a_3b_1+cb_2-a_1b_3)}{2b_2T_1^2}$	$\frac{(a_3b_3-a_4b_2)(-a_3b_1+cb_2-a_1b_3)}{2b_3T_1^2}$	$\frac{a_0b_2b_3-T_2T_3}{b_1b_2b_3}$
L_1	0	$\frac{T_3^2+a_0b_2^2}{b_1b_2^2}$	$\frac{T_3^2+a_0b_2^2}{b_1b_2^2}$	0	$\frac{T_2^2+a_0b_3^2}{b_1b_3^2}$	$\frac{cb_3-a_4b_1}{b_3^2}$	0	$\frac{\mp U_1(a_1b_3-cb_2)+c(a_3b_2-2db_3)+a_1(a_3b_3-2a_4b_2)}{2T_1^2}$	$\frac{c(a_3b_2-db_3)-a_4(a_1b_2-db_1)}{T_1^2}$	$\frac{d(a_4b_1-cb_3)-a_1(a_4b_2-a_3b_3)}{T_1^2}$	0
L_2	$\frac{U_1(a_1b_3-cb_2)+c(a_3b_2-2db_3)+a_1(a_3b_3-2a_4b_2)}{2T_1^2}$	$\frac{b_1T_1^2}{b_2^2b_3^2}$	$\frac{b_1(a_3b_2-db_3)^2}{b_2^2T_1^2}$	$\frac{d(a_4b_1-cb_3)+a_1(a_3b_3-a_4b_2)}{T_1^2}$	$\frac{b_1T_1^2}{b_2^2b_3^2}$	$\frac{b_1(a_3b_3-a_4b_2)2}{b_3^2T_1^2}$	$\frac{c(a_3b_2-db_3)-a_4(a_1b_2-db_1)}{T_1^2}$	$\frac{b_1U_1^2}{T_1^2}$	$\frac{b_1(a_3b_2-db_3)^2}{b_2^2T_1^2}$	$\frac{b_1(a_4b_2-a_3b_3)2}{b_3^2T_1^2}$	$\frac{a_0b_3^2+T_2^2}{b_1b_3^2}$
L_3	$\frac{-U_1(a_1b_3-cb_2)+c(a_3b_2-2db_3)+a_1(a_3b_3-2a_4b_2)}{2T_1^2}$	0	0	$\frac{cb_3-a_4b_1}{b_3^2}$	0	0	$\frac{a_1b_2-db_1}{b_2^2}$	0	0	0	$\frac{a_0b_2^2+T_3^2}{b_1b_2^2}$
R_1	0	$\frac{d}{b_2}$	$\frac{d}{b_2}$	0	$\frac{d}{b_2}$	0	$\frac{d}{b_2}$	0	$\frac{d}{b_2}$	0	$\frac{d}{b_2}$
R_2	$\frac{b_2U_1^2-(a_3b_2-2db_3)U_1}{2T_1^2}$	$\frac{T_1^2}{b_2b_3^2}$	$\frac{(a_3b_2-db_3)^2}{b_2T_1^2}$	$\frac{d(a_3b_3-a_4b_2)}{T_1^2}$	0	$\frac{d(a_3b_3-a_4b_2)}{T_1^2}$	$\frac{(a_3b_2-db_3)^2}{b_2T_1^2}$	$\frac{b_2U_1^2\mp(a_3b_2-2db_3)U_1}{2T_1^2}$	0	$\frac{a_3b_3-a_4b_2}{b_3^2}$	$\frac{T_1^2}{b_2b_3^2}$
R_3	$\frac{b_2U_1^2+(a_3b_2-2db_3)U_1}{2T_1^2}$	0	0	$\frac{a_3b_3-a_4b_2}{b_3^2}$	$\frac{T_1^2}{b_2b_3^2}$	$\frac{a_3b_3-a_4b_2}{b_3^2}$	0	$\frac{b_2U_1^2\pm(a_3b_2-2db_3)U_1}{2T_1^2}$	$\frac{(a_3b_2-db_3)^2}{b_2T_1^2}$	$\frac{d(a_3b_3-a_4b_2)}{T_1^2}$	0
C_1	∞	$\frac{b_3}{a_4}$	∞	$\frac{b_3}{a_4}$	$\frac{b_3}{a_4}$	$\frac{b_3}{a_4}$	∞	∞	∞	$\frac{b_3}{a_4}$	$\frac{b_3}{a_4}$
C_2	$\frac{2T_1^2}{b_3U_1^2+(a_3b_3-2a_4b_2)U_1}$	∞	$\frac{T_1^2}{a_4(a_3b_2-db_3)}$	$\frac{b_3T_1^2}{(a_3b_3-a_4b_2)^2}$	$\frac{b_2^2b_3}{T_1^2}$	$\frac{b_3T_1^2}{(a_3b_3-a_4b_2)^2}$	$\frac{T_1^2}{a_4(a_3b_2-db_3)}$	$\frac{2T_1^2}{b_3U_1^2\pm(a_3b_3-2a_4b_2)U_1}$	$\frac{b_2^2}{a_3b_2-db_3}$	∞	∞
C_3	$\frac{2T_1^2}{b_3U_1^2-(a_3b_3-2a_4b_2)U_1}$	$\frac{b_2^2b_3}{T_1^2}$	$\frac{b_2^2}{a_3b_2-db_3}$	∞	∞	∞	$\frac{b_2^2}{a_3b_2-db_3}$	$\frac{2T_1^2}{b_3U_1^2\mp(a_3b_3-2a_4b_2)U_1}$	$\frac{T_1^2}{a_4(a_3b_2-db_3)}$	$\frac{b_3T_1^2}{(a_3b_3-a_4b_2)^2}$	$\frac{b_2^2b_3}{T_1^2}$

$$S=\frac{a_0\lambda^4+a_1\lambda^3+a_2\lambda^2+a_3\lambda+a_4}{b_1\lambda^3+b_2\lambda^2+b_3\lambda},\quad \lambda=ip,$$

$$d^2(b_2^2-4b_1b_3)-2d(2a_4b_1^2+a_2b_2^2+2a_0b_3^2-a_3b_1b_2-2a_2b_1b_3-a_1b_2b_3)+\left[a_3^2b_1^2+(a_2^2-4a_0a_4)b_2^2+a_1^2b_3^2-2(a_2a_3-2a_1a_4)b_1b_2-2a_1a_3b_1b_3-2(a_1a_2-2a_0a_3)b_2b_3\right]=0,$$

$c=a_2-d,$

$$T_1=\pm\sqrt{-a_4b_2^2+a_3b_2b_3-db_3^2},\qquad U_1=\sqrt{a_3^2-4a_4d},$$

$$T_2=\pm\sqrt{-a_0b_3^2+cb_3b_1-a_4b_1^2},\qquad U_2=\sqrt{c^2-4a_0a_4},$$

$$T_3=\pm\sqrt{-db_1^2+a_1b_1b_2-a_0b_2^2},\qquad U_3=\sqrt{a_1^2-4a_0d}.$$

with signs chosen so that

$$b_1T_1+b_2T_2+b_3T_3=0,$$

Table II

Networks	12	13	14	15	16	17	18	19	20	21	22	23
L_1	$\frac{a_0T_1^2-dT_2^2-a_4T_3^2-a_3T_2T_3}{b_1T_1^2}$	$\frac{a_0b_2b_3-T_2T_3}{b_1b_2^2b_3}$	$\frac{(a_1b_2-db_1)T_1+(a_3b_2-db_3)T_3}{b_2^2T_1}$	$\frac{(cb_3-a_4b_1)T_1+(a_3b_3-a_4b_2)T_2}{b_3^2T_1}$	0	$\frac{a_0}{b_1}$	$\frac{a_0}{b_1}$	0	0	$\frac{a_0}{b_1}$	$\frac{a_0}{b_1}$	0
L_2	$\frac{b_1U_1^2+(a_1b_3-cb_2)U_1}{2T_1^2}$	$-\frac{T_1T_2}{b_2b_3^2}$	$\frac{(a_3b_2-db_3)(a_3b_1+cb_2-a_1b_3)}{2b_2T_1^2}$	$\frac{(a_3b_3-a_4b_2)(a_3b_1-cb_2+a_1b_3)}{2b_3T_1^2}$	$\frac{b_1U_2^2-(cb_1-2a_0b_3)U_2}{2T_2^2}$	$\frac{T_2^2}{b_1b_3^2}$	$\frac{(cb_1-a_0b_3)^2}{b_1T_2^2}$	$\frac{a_0(cb_3-a_4b_1)}{T_2^2}$	$\frac{b_1U_3^2+(a_1b_1-2a_0b_2)U_3}{2T_3^2}$	$\frac{T_3^2}{b_1b_2^2}$	$\frac{(a_1b_1-a_0b_2)^2}{b_1T_3^2}$	$\frac{a_0(a_1b_2-db_1)}{T_3^2}$
L_3	$\frac{b_1U_1^2-(a_1b_3-cb_2)U_1}{2T_1^2}$	$-\frac{T_1T_3}{b_2^2b_3}$	$\frac{(db_3-a_3b_2)T_3}{b_2^2T_1}$	$\frac{(a_4b_2-a_3b_3)T_2}{b_3^2T_1}$	$\frac{b_1U_2^2+(cb_1-2a_0b_3)U_2}{2T_2^2}$	0	0	$\frac{cb_3-a_4b_1}{b_3^2}$	$\frac{b_1U_3^2-(a_1b_1-2a_0b_2)U_3}{2T_3^2}$	0	0	$\frac{a_1b_2-db_1}{b_2^2}$
R_1	0	$\frac{d}{b_2}$	$\frac{d}{b_2}$	0	$\frac{dT_2^2-a_4T_3^2-a_0T_1^2-cT_1T_3}{b_2T_2^2}$	$\frac{db_1b_3-T_1T_3}{b_1b_2b_3}$	$\frac{(a_1b_1-a_0b_2)T_2+(cb_1-a_0b_3)T_3}{b_1^2T_2}$	$\frac{(cb_3-a_4b_1)T_1+(a_3b_3-a_4b_2)T_2}{b_3^2T_2}$	0	$\frac{d}{b_2}$	0	$\frac{d}{b_2}$
R_2	$\frac{b_2U_1^2-(a_3b_2-2db_3)U_1}{2T_1^2}$	$\frac{T_1^2}{b_2b_3^2}$	$\frac{(a_3b_2-db_3)^2}{b_2T_1^2}$	$\frac{d(a_3b_3-a_4b_2)}{T_1^2}$	$\frac{b_2U_2^2+(a_1b_3-a_3b_1)U_2}{2T_2^2}$	$-\frac{T_1T_2}{b_1b_3^2}$	$\frac{(cb_1-a_0b_3)(cb_2+a_3b_1-a_1b_3)}{2b_1T_2^2}$	$\frac{(cb_3-a_4b_1)(-a_3b_1+cb_2+a_1b_3)}{2b_3T_2^2}$	$\frac{b_2U_3^2-(a_1b_2-2db_1)U_3}{2T_3^2}$	0	$\frac{d(a_1b_1-a_0b_2)}{T_3^2}$	$\frac{(a_1b_2-db_1)^2}{b_2T_3^2}$
R_3	$\frac{b_2U_1^2+(a_3b_2-2db_3)U_1}{2T_1^2}$	0	0	$\frac{a_3b_3-a_4b_2}{b_3^2}$	$\frac{b_2U_2^2-(a_1b_3-a_3b_1)U_2}{2T_2^2}$	$-\frac{T_2T_3}{b_1^2b_3}$	$\frac{(a_0b_3-cb_1)T_3}{b_1^2T_2}$	$\frac{(a_4b_1-cb_3)T_1}{b_3^2T_2}$	$\frac{b_2U_3^2+(a_1b_2-2db_1)U_3}{2T_3^2}$	$\frac{T_3^2}{b_1^2b_2}$	$\frac{a_1b_1-a_0b_2}{b_1^2}$	0
C_1	∞	$\frac{b_3}{a_4}$	∞	$\frac{b_3}{a_4}$	∞	$\frac{b_3}{a_4}$	∞	$\frac{b_3}{a_4}$	$\frac{b_3T_3^2}{a_4T_3^2-dT_2^2-a_0T_1^2-a_1T_1T_2}$	$\frac{b_1b_2b_3}{a_4b_1b_2-T_1T_2}$	$\frac{b_1^2T_3}{(cb_1-a_0b_3)T_3+(a_1b_1-a_0b_2)T_2}$	$\frac{b_2^2T_3}{(a_3b_2-db_3)T_3+(a_1b_2-db_1)T_1}$
C_2	$\frac{2T_1^2}{b_3U_1^2+(a_3b_3-2a_4b_2)U_1}$	∞	$\frac{T_1^2}{a_4(a_3b_2-db_3)}$	$\frac{b_3T_1^2}{(a_3b_3-a_4b_2)^2}$	$\frac{2T_2^2}{b_3U_2^2+(cb_3-2a_4b_1)U_2}$	∞	$\frac{T_2^2}{a_4(cb_1-a_0b_3)}$	$\frac{b_3T_2^2}{(cb_3-a_4b_1)^2}$	$\frac{2T_3^2}{b_3U_3^2+(a_3b_1-cb_2)U_3}$	$-\frac{b_1b_2^2}{T_1T_3}$	$\frac{2b_1T_3^2}{(a_1b_1-a_0b_2)(a_1b_3-cb_2+a_3b_1)}$	$\frac{2b_2T_3^2}{(a_1b_2-db_1)(a_1b_3+cb_2-a_3b_1}$
C_3	$\frac{2T_1^2}{b_3U_1^2-(a_3b_3-2a_4b_2)U_1}$	$\frac{b_2^2b_3}{T_1^2}$	$\frac{b_2^2}{a_3b_2^2-db_3}$	∞	$\frac{2T_2^2}{b_3U_2^2-(cb_3-2a_4b_1)U_2}$	$\frac{b_1^2b_3}{T_2^2}$	$\frac{b_1^2}{cb_1-a_0b_3}$	∞	$\frac{2T_3^2}{b_3U_3^2-(a_3b_1-cb_2)U_3}$	$-\frac{b_1^2b_2}{T_2T_3}$	$\frac{b_1^2T_3}{(a_0b_2-a_1b_1)T_2}$	$\frac{b_2^2T_3}{(db_1-a_1b_2)T_1}$

for the computation of the elements of this general network can be stated in the form of the following theorem

Theorem IV. Any driving-point impedance satisfying the necessary and sufficient conditions, as stated in Theorem I, can be realized

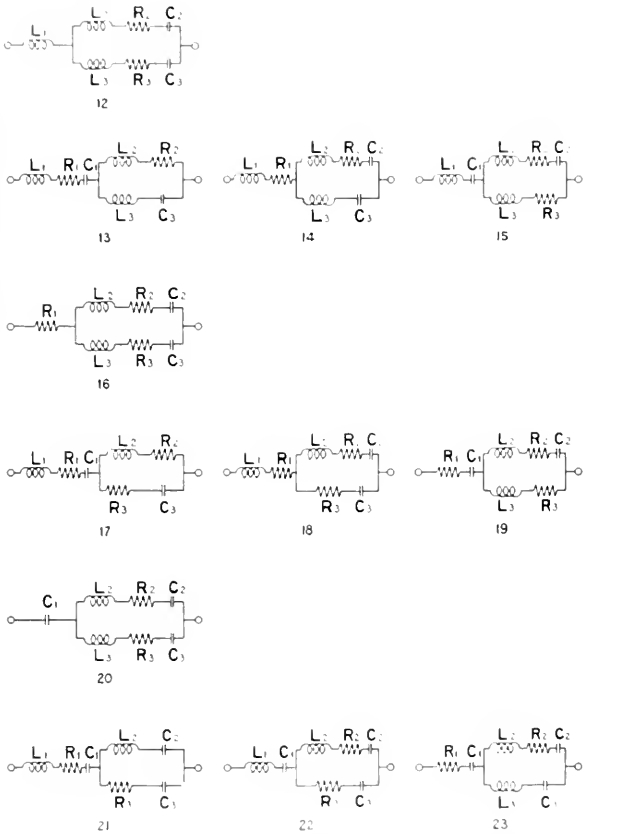


Fig. 2. Networks without mutual inductance realizing any driving-point impedance of a two-mesh circuit consisting of resistances, capacities, and self-inductances

by any network of the form of Fig. 7, provided the elements of the network satisfy the following relations:

$$L_1'L_2'+L_1'L_3'+L_2'L_3'=a_0k^2, \tag{7}$$

$$R_1R_2+R_1R_3+R_2R_3=dk^2, \tag{8}$$

$$D_1D_2+D_1D_3+D_2D_3=a_1k^2, \tag{9}$$

$$L_3'+L_3'=b_1k^2, \tag{10}$$

$$R_2+R_3=b_2k^2, \tag{11}$$

$$D_2+D_3=b_3k^2, \tag{12}$$

$$R_2D_3-R_3D_2=\pm k^3(-a_1b_2^2+a_3b_2b_3-db_3^2)^{1/2}, \tag{13}$$

$$D_2L_3'-D_3L_2'=\pm k^3[-a_0b_3^2+(a_2-d)b_3b_1-a_1b_1^2]^{1/2}, \tag{14}$$

$$L_2'R_3-L_3'R_2=\pm k^3(-db_1^2+a_1b_1b_2-a_0b_2^2)^{1/2}, \tag{15}$$

where

$$D_1=C_1^{-1},\ D_2=C_2^{-1},\ D_3=C_3^{-1}, \tag{16}$$

and

$$L_1'=L_1+M_{12}+M_{13}+M_{23}, \tag{17}$$

$$L_2'=L_2+M_{12}-M_{13}-M_{23}, \tag{18}$$

$$L_3'=L_3-M_{12}+M_{13}-M_{23}, \tag{19}$$

the positive directions in Fig. 7 all being assigned arbitrarily to the right. The signs of (13) (15) are chosen so as to satisfy the identity

$$\begin{aligned}(R_2D_3-R_3D_2)(L_2'+L_3')+(D_2L_3'-D_3L_2')(R_2+R_3)\\+(L_2'R_3-L_3'R_2)(D_2+D_3)=0.\end{aligned} \tag{20}$$

The value of d is given by equation (3), which may be written in the form

$$\begin{aligned}d^2(b_2^2-4b_1b_3)-2d(2a_1b_1^2+a_2b_2^2+2a_0b_3^2-a_3b_1b_2-2a_2b_1b_3-a_1b_2b_3)\\+[a_1^2b_1^2+(a_2^2-4a_0a_1)b_2^2+a_1^2b_3^2-2(a_2a_3-2a_1a_1)b_1b_2-2a_1a_3b_1b_3\\-2(a_1a_2-2a_0a_3)b_2b_3]=0.\end{aligned} \tag{21}$$

The parameter k may have any real value other than zero.

In these formulas the value of k is independent of the impedance, but can be chosen so as to give particular forms of the network. If the necessary and sufficient conditions as stated by Theorem I are satisfied, the values of the elements given by these formulas are positive or zero, and the values of the inductances satisfy the usual restrictions. The formulas of Tables I and II, for example, can all be computed by means of Theorem IV.

2. The Driving-Point Impedance of a Two-Mesh Circuit

Previous investigations of the two-mesh circuit have been directed, for the most part, toward the determination of the free periods (reso-

nant frequencies and associated damping constants) of the circuit from the known values of the elements. This problem is intimately related to the determination of the driving-point impedance of the circuit, since the free periods of the circuit can be found by setting the driving-point impedance in any one mesh equal to zero.[4] By this method the free periods are found as the roots of an equation of the fourth degree,[5] the exact solution of which involves, in general, cumbersome formulas. In order to obtain formulas which are better adapted to numerical computation, various approximations are usually made.[6]

This electrical problem of the free oscillations of a circuit is formally the same as the dynamical problem of the small oscillations of a system about a position of equilibrium. The determination of the free periods of a circuit can be made directly from the solution of this dynamical problem.[7]

The first part of this paper treats a much more general problem than the determination of the driving-point impedance of a particular circuit from the given values of the elements, namely, the determination of the entire range of possibilities, together with the inherent limitations, of such an impedance. The method employed is to find the general form of the impedance as a function of the time coefficient, and then to investigate the restrictions which must be satisfied by a function of this character in order that it may represent an impedance realizable by means of a circuit consisting of resistances, capacities, and inductances. In the present paper, this investigation is limited to the driving-point impedance of a two-mesh circuit; the driving-point impedance of an n-mesh circuit will be treated in a future paper.

The driving-point impedance of any circuit containing no resistances has been investigated in a previous paper,[8] where it has been shown that any such impedance is a pure reactance with a number of resonant and anti-resonant frequencies which alternate with each other, and

[4] G. A. Campbell, *Transactions of the A. I. E. E.*, 30, 1911, pages 873–909.

[5] An exhaustive discussion of this fourth degree equation has been given by J. Sommer, *Annalen der Physik*, fourth series, 58, 1919, pages 375–392.

[6] For typical methods of solution see the papers of L. Cohen, *Bulletin of the Bureau of Standards*, 5, 1908–9, pages 511–541; B. Macku, *Jahrbuch der drahtlosen Telegraphie und Telephonie*, 2, 1909, pages 251–293; V. Bush, *Proceedings of the I. R. E.*, 5, 1917, pages 363–382.

[7] Representative investigations of this dynamical problem are those of Lord Rayleigh, *Proceedings of the London Mathematical Society*, 4, 1873, pages 357–368, *Philosophical Magazine*, fifth series, 21, 1886, pages 369–381, and sixth series, 3, 1902, pages 97–117 ("Scientific Papers," I, 170–181, II, 475–485, and V, 8–26); E. J. Routh, "Advanced Rigid Dynamics," sixth edition, 1905, pages 232–243; A. G. Webster, "Dynamics," second edition, 1912, pages 157–164.

[8] R. M. Foster, *Bell System Technical Journal*, 3, 1924, pages 259–267.

that any such impedance may be realized by a network consisting of a number of simple resonant elements (inductance and capacity in series) in parallel or a number of simple anti-resonant elements (inductance and capacity in parallel) in series.

With resistances added to the circuit, the impedance is, in general, complex; that is, it has both resistance and reactance components. For a two-mesh circuit the impedance is expressed as a function of the time coefficient by Theorem I.

Formula (1) gives the driving-point impedance of a two-mesh circuit for any electrical oscillation of the form $e^{\lambda t}$, where the time coefficient λ may have any value, real or complex. The time coefficients for the free oscillations of the circuit with the driving branch closed are the roots of the numerator $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$, as given by (1a); the free periods of the circuit with the driving branch opened are the roots of the denominator (β_2, β_3), that is, the poles of the impedance function. For a complex value of the time coefficient, $\lambda = \lambda_1 + i\lambda_2$, λ_1 is the damping factor and λ_2 is the frequency multiplied by 2π.

The two forms of formula (1) are equivalent, but each has its special advantages. Sometimes one, sometimes the other, form is more convenient; they will be used interchangeably throughout the paper.

Formula (1a) gives the impedance directly in terms of the roots and poles. Formula (1b) gives the impedance in terms of the symmetric functions of the roots and poles, with the addition of an arbitrary factor. Thus, without changing the impedance, all the coefficients of the numerator and denominator of (1b) may be multiplied by the same constant factor having any value other than zero. Formulas stated in terms of the coefficients of (1b) are in homogeneous and symmetrical form, and have the added advantage of involving real quantities only.

The special case of one root equal to zero is obtained by setting $\alpha_1 = 0$ in (1a) and $a_4 = 0$ in (1b). For one root infinite, however, in (1a) it is necessary to set $\alpha_1 = \infty$ and $H = 0$, with the provision that $H\alpha_1$ be finite; whereas in (1b) it is simply necessary to set $a_0 = 0$.

It is sometimes convenient to add the notation $\beta_1 = 0$ and $\beta_4 = \infty$, corresponding to the poles at zero and infinity. In formula (1b) the corresponding addition to the notation consists of the coefficients $b_0 = 0$ and $b_4 = 0$.

By the general restrictions (2) the constant H is positive or zero, and the roots and poles are arranged in three pairs, (α_1, α_2), (α_3, α_4), and (β_2, β_3), each pair being the roots of a quadratic equation with positive real coefficients. Thus each pair of the roots and poles is

either a pair of conjugate complex quantities or a pair of real quantities, with the added provision that the real part of each root and pole is negative or zero.

Stated in terms of (1b), these general restrictions (2) require all the coefficients to be real and to have the same sign. Throughout this paper these signs will always be taken positive; thus all the a's and b's are positive or zero. In order to provide that the real part of each root be negative or zero, the coefficients of the numerator must satisfy the additional requirement

$$-a_4a_1^2+a_1a_2a_3-a_0a_3^2\geq 0, \tag{22}$$

and also

$$a_2^2-4a_0a_4\geq 0. \tag{23}$$

The second condition (23) is satisfied automatically by virtue of the first condition (22), unless both a_1 and a_3 are zero; in that case (23) is required. These are precisely the necessary and sufficient conditions that the numerator of (1b) be factorable into two real quadratic factors with positive coefficients.

In addition to the general restrictions (2) upon the individual roots and poles, there are certain additional conditions which must be satisfied by all the roots and poles together. These conditions are more conveniently stated in terms of the coefficients by prescribing a certain domain of values of the eight coefficients (a_0, a_1, a_2, a_3, a_4, b_1, b_2, b_3) such that the coefficients of any driving-point impedance of a two-mesh circuit lie in this domain, and, conversely, any set of values in this domain can be realized as the coefficients of a driving-point impedance of a two-mesh circuit.

By a realizable circuit is understood a circuit consisting of resistances, capacities, and self-inductances, with positive or zero values, together with mutual inductances with values such that every principal minor of the determinant of the inductances is positive or zero. In the case of two self-inductances with mutual inductance between them, this reduces to the well known condition $L_1L_2-M^2\geq 0$.

The domain is defined analytically by formulas (3)–(6), in terms of a parameter d. This parameter is intimately related to the resistances in the circuit, as will be shown later. In order that this domain may contain real values, the following relation must be satisfied:

$$-d^3+2a_2d^2-(a_1a_3+a_2^2-4a_0a_4)d+(-a_4a_1^2+a_1a_2a_3-a_0a_3^2)\geq 0, \tag{24}$$

or in equivalent form,

$$\begin{aligned}-[d-a_0(\alpha_1+\alpha_2)(\alpha_3+\alpha_4)][d-a_0(\alpha_1+\alpha_3)(\alpha_2+\alpha_4)]\\ [d-a_0(\alpha_1+\alpha_4)(\alpha_2+\alpha_3)]\geq 0.\end{aligned} \tag{25}$$

Provided there is one pair of conjugate complex roots of the numerator of the impedance, α_1 and α_2, the value of d is restricted to the range from zero to the smallest real root of (24), that is,

$$0 \leq d \leq a_0(\alpha_1+\alpha_2)(\alpha_3+\alpha_4). \tag{26}$$

In the case of four real roots, $\alpha_1 \geq \alpha_2 \geq \alpha_3 \geq \alpha_4$, the parameter d is restricted to the values

$$\begin{gathered} 0 \leq d \leq a_0(\alpha_1+\alpha_2)(\alpha_3+\alpha_4), \\ a_0(\alpha_1+\alpha_3)(\alpha_2+\alpha_4) \leq d \leq a_0(\alpha_1+\alpha_4)(\alpha_2+\alpha_3). \end{gathered} \tag{27}$$

Thus there are, in general, two distinct ranges for the value of d in this case. The corresponding domain of values of the roots and poles consists of two non-connected regions, so that it is impossible to pass by a continuous transformation from a set of values in one region to a set in the other.

Formulas (3) (6) are symmetrical in three different respects, since they remain unaltered upon interchanging certain pairs of elements, which may be any one of the three following sets:

$$\begin{aligned} &\text{(a)}\ \ b_1 \text{ and } b_2,\ a_0 \text{ and } d,\ a_3 \text{ and } (a_2-d), \\ &\text{(b)}\ \ b_1 \text{ and } b_3,\ a_0 \text{ and } a_4,\ a_1 \text{ and } a_3, \\ &\text{(c)}\ \ b_2 \text{ and } b_3,\ a_4 \text{ and } d,\ a_1 \text{ and } (a_2-d). \end{aligned} \tag{28}$$

These three sets correspond to interchanging resistances and inductances, inductances and capacities, and resistances and capacities, respectively.

Since d is always positive or zero, formulas (4) (6) lead to simple necessary conditions, namely,

$$a_3 b_3 - a_4 b_2 \geq 0, \tag{29}$$

$$-a_4 b_1^2 + a_2 b_1 b_3 - a_0 b_3^2 \geq 0, \tag{30}$$

$$a_1 b_1 - a_0 b_2 \geq 0. \tag{31}$$

The first and third of these conditions are conveniently interpreted in terms of the roots and poles: the sum of the reciprocals of the poles is algebraically greater than or equal to the sum of the reciprocals of the roots; and the sum of the poles is algebraically greater than or equal to the sum of the roots.

3. Domain of Poles for Assigned Roots

The conditions (2) (6) define a domain of values for the roots and poles without distinguishing in any way those roots and poles which may be chosen independently. For many purposes it is convenient

to specialize the problem to the extent of assigning definite values to the roots, subject, of course, to the restrictions (2), and then to investigate the domain of the poles which can be associated with these assigned roots.

For the mathematical analysis of the problem it is convenient to assign values of the coefficients $a_0, \ldots, a_4$, subject to the restrictions stated in the preceding section, and then to plot the domain for the coefficients b_1, b_2, b_3, treating the latter as homogeneous coordinates[4] in the plane, with $x = b_2 - b_1$ and $y = b_3 - b_1$.

With this method of representation, equation (3) is, for any fixed value of d, the equation of a conic. Considering d as a variable parameter, (3) represents a one-parameter family of conics. Each curve of this family is tangent to the four lines

$$\alpha_j^2 b_1 + \alpha_j b_2 + b_3 = 0, \qquad (j = 1, 2, 3, 4). \tag{32}$$

These lines are real lines in the plane if, and only if, the corresponding roots are real. They are all tangent to the parabola

$$b_2^2 - 4 b_1 b_3 = 0, \tag{33}$$

which is the limiting case of the conic (3) as d becomes infinite. This parabola is a critical curve for the poles; every point in the plane above the parabola corresponds to a pair of conjugate complex poles, every point below the curve to a pair of real and distinct poles, and every point on the curve to a pair of real and equal poles.

The complete family of conics, that is, the set of curves for all real values of d, might be defined as the family of conics tangent to these four lines, which are the four lines tangent to the critical parabola (33) corresponding to the four roots of the impedance.

Not all the curves of this family lie in the domain of poles, however, since the conditions (4) (6) must also be satisfied. For any fixed value of d, each of the three equations (4) (6) is a degenerate conic, that is, a pair of straight lines. The six lines defined by these conditions are all tangent to the conic (3) corresponding to this same value of d. The inequalities (4) (6) thus demand, in general, that the domain of poles lie within the area bounded by these six lines. Thus only those conics of the family (3) which are real ellipses, or their limiting cases, lie within the domain.

The condition that the conic (3) be an ellipse is precisely the necessary restriction on the value of d already stated, formula (21). Ellipses are obtained for all negative values of d, but these are not in the

[4] For some purposes the other choices of x and y might be used, this choice is more convenient here inasmuch as $-x$ is the sum and y the product of the poles.

domain, since by the conditions of the electrical problem d must be positive or zero. Ellipses for values of d from zero up to the smallest real root of the equation (24) are in the domain. If the roots of the impedance are all complex, equation (24) has three real roots, and thus there is a range of values of d from the second to the third root, arranged in the order of magnitude, for which the curves are ellipses, but these ellipses are imaginary, that is, there are no real points on them; thus there is only the one range of d which gives points in the domain. If two roots of the impedance are real and two complex, equation (24) has only the one real root, and thus there is only the one range of d. If all four roots of the impedance are real, however, equation (24) has again three real roots, and both ranges of d give real ellipses. In this case the two sets of ellipses are separate and distinct.

For the limiting values of d, that is, for the roots of equation (24), the corresponding conic (3) degenerates into a pair of coincident straight lines. Only those segments of these lines which satisfy the corresponding inequalities (4) (6) are in the domain. Such segments are the limiting cases of the real ellipses for values of d above or below the critical values, as the case may be.

The domain of poles, plotted in terms of the coefficients in the manner described, consists of that domain covered by these real ellipses for $d \geq 0$, a domain bounded by the envelope of the curves. The envelope consists of the conic for $d = 0$ and the four lines (32). For the case of four complex roots of the impedance, therefore, the domain consists simply of the region bounded by the ellipse (3) for $d = 0$. For two complex and two real roots, the domain consists of the region bounded by the ellipse with the addition of the corner bounded by the ellipse and the two tangent lines to the ellipse corresponding to the two real roots. For four real roots, the domain consists of the region bounded by the ellipse together with the two corners bounded by the ellipse and the tangent lines, one by the two lines corresponding to the two smallest roots and the other the two largest roots; and a second region consisting of the quadrilateral bounded by the four tangent lines.

All points in the domain lying on or above the critical parabola lie on a single curve of the family of conics composing the domain, points below the parabola on two curves of the family. The corner regions and the quadrilateral are entirely below the critical parabola. Where there is a corner region, the ellipse goes below the parabola, otherwise not.

The foregoing discussion has all been for the general case of un-

restricted roots. For special cases of zero, pure imaginary, or infinite roots, the corresponding domains are the limiting cases of the general domain, described above. Such limiting cases may reduce to a single segment or to a region bounded in part by the line at infinity. The homogeneous coordinates employed are very useful in dealing with these special cases.

4. Figures Illustrating the Domain of Poles

The preceding section presented a discussion of the domain of the poles associated with any four assigned roots, the domain being plotted in terms of the coefficients of the denominator of the impedance, that is, in terms of symmetric functions of the poles. In order to show the mutual relations between the actual values of the roots and the poles, it is convenient to plot, in the upper half of the complex plane, the domain of one pole, the other pole being its conjugate. This provides a complete representation for the case of complex poles. In order to include the domain of real poles, an auxiliary graph can be provided to indicate pairs of points on the real axis.

The mathematical analysis for this form of representation can be obtained from that of the preceding section by substituting $\beta_2+\beta_3=-b_2\ b_1$ and $\beta_2\beta_3=b_3\ b_1$. For complex poles, $\beta_2=u+iv$ and $\beta_3=u-iv$, this transformation from the x, y plane to the u, v plane is simply $2u=-x$ and $u^2+v^2=y$. Thus a conic in the x, y plane becomes, in general, a curve of the fourth degree in the u, v plane. The analysis of the curves obtained in the u, v plane is not so simple as in the other plane, but there is a decided advantage in the interpretation of the results in this plane, since the coordinate u, the real part of the pole, corresponds to the damping factor, and the coordinate v, the imaginary part of the pole, corresponds to the frequency factor.

In the complex plane, the necessary conditions (29)–(31) require the domain of complex poles to lie entirely within the region bounded by the vertical axis, a vertical line to the left of the axis, two circles about the origin as center, and a circle through the origin with its center on the real axis. Furthermore, the boundary curve of the domain must be tangent to each of these lines and circles, since the corresponding conic (3) for $d=0$ is tangent to the corresponding lines (4)–(6) for $d=0$.

For the special case of one root a positive pure imaginary, the second root being its conjugate, the domain in the upper half of the complex plane reduces merely to the points on an arc of a circle with its center on the real axis. If the third root is complex with a positive imaginary part, the fourth root being its conjugate, the domain

is the circular arc extending from the first root to the third root. For a pure imaginary value of the third root the radius of the circle becomes infinite, and the domain is the segment of the vertical axis between the first and third roots. This is precisely the result already obtained for the resistanceless circuit.

For the limiting case of the third root real, with the fourth root equal to it, the domain is the circular arc extending from the root on the imaginary axis to the double root on the real axis. When the third and fourth roots are real and distinct, the domain is the circular arc from the first root to the point on the real axis midway between the two real roots. The complete domain also includes real poles in the segment between the two real roots, equally spaced about the midpoint of the segment.

This case of one pair of roots on the axis of imaginaries is illustrated by Fig. 3a, with the first root fixed at the point a, and the third root lying on any one of the family of circular arcs drawn through a, the fourth root being its conjugate; or the third and fourth roots lying on the real axis equally spaced about the end-point of one of the arcs.

Starting with one pair of roots on the axis of imaginaries, it is interesting to investigate the changes made in the domain by moving this pair of roots off the axis. The domain broadens out into a region lying about the circular arc, as shown by Fig. 3b for four typical cases. The first case is for the third root also near the axis ($\alpha_1 = -0.5 + i3$, $\alpha_3 = -0.5 + i9$); and the second case is for the third root some distance from the axis ($\alpha_1 = -0.1 + i3$, $\alpha_3 = -5 + i8$). The third section of Fig. 3b shows the domain when the third and fourth roots are real and equal ($\alpha_1 = -0.1 + i3$, $\alpha_3 = \alpha_4 = -9$); in this case the region has a cusp at this double root. The fourth section shows the domain of complex poles when the third and fourth roots are real and distinct ($\alpha_1 = -0.1 + i3$, $\alpha_3 = -6$, $\alpha_4 = -10$); in this case the region of complex poles terminates along a segment of the real axis lying in the interval between the two real roots, there is also a domain of real poles which is not shown.

It is interesting to note that, when both pairs of roots are near the axis of imaginaries, that is, for small damping, the frequency factor of the pole may always be taken outside the range of the frequency factors of the roots; whereas for zero damping the pole must lie between the roots, as noted above.

Fig. 3c shows the domain of the poles for two pairs of equal roots. If the first and third roots are equal, the second and fourth roots being their conjugates and thus also equal, the domain is bounded

by a circle tangent to the vertical axis with its center vertically above the double root. If, for example, the double root describes a circle about the origin through the point a on the vertical axis, the corresponding circle is tangent to the vertical axis at a. Thus in Fig. 3(c),

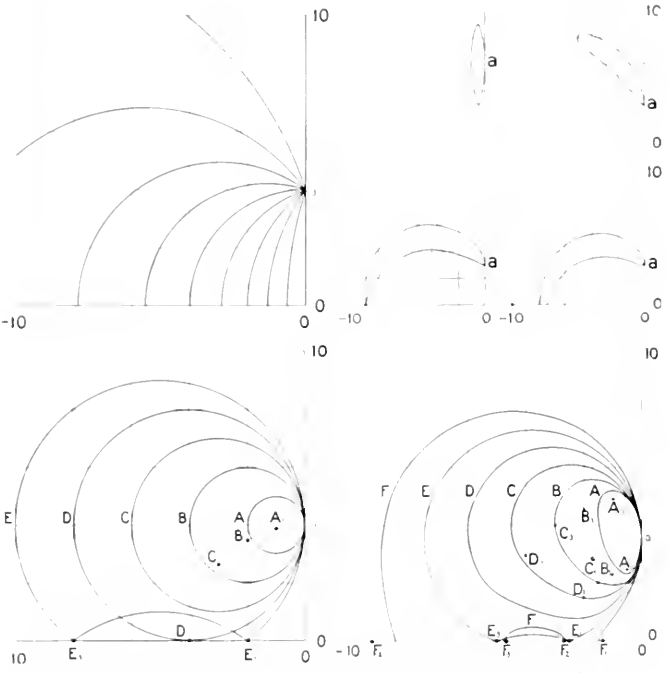


Fig. 3. Domain of the poles of the driving point impedance of a two mesh circuit with (a) one pair of roots on the axis of imaginaries, (b) one pair of roots near the axis of imaginaries, (c) two pairs of equal roots, and (d) two pairs of roots with equal angles.

for double roots at A, B, C, the corresponding domain is bounded by the circles A, B, C, respectively. The centers of these circles are all on the horizontal line through a, and the double roots are selected so as to space the centers uniformly. If all four roots are real and equal, the domain is bounded by a circle D tangent to the vertical axis at a and to the horizontal axis at this fourfold root D. If the

roots are all real and equal in pairs the domain is bounded by a circle E_1 tangent to the vertical axis and passing through the two double roots, E_1 and E_3, and by the reflection of this circle in the real axis. Thus the domain has cusps at the double roots. For two pairs of

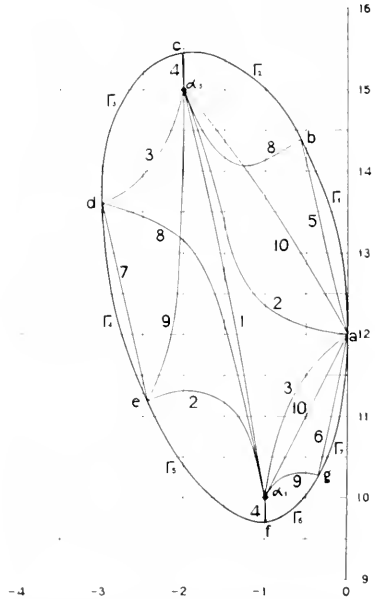


Fig. 4—Domain of the poles of the driving-point impedance of a two-mesh circuit with two pairs of complex roots, showing the portions of the domain realizable by each network listed in Table III.

equal roots, whether real or complex, the distance Oa is the geometrical mean value of all four roots.

Another kind of special case is shown by Fig. 3d, the case of two pairs of roots with equal angles. The first and third roots are on a line with the origin, so that the second and fourth roots, being their conjugates, are also on a line with the origin. Fig. 3d shows the boundary curves ($A \ldots E$) for five sets of roots ($A_1, A_3 \ldots E_1, E_3$) satisfying these conditions and with the same absolute values of the roots in each set, so that the roots lie on two circles about the origin.

The fifth set of roots (E_1, E_4) has a domain of the same type as the corresponding set of roots on Fig. 3c, since this set, being on the real axis, is a double set. The sixth curve F is the boundary of the domain for four real roots so chosen that $F_1F_3 = E_1^2$ and $F_2F_4 = E_1^2$. This is the same type of domain as will be described later under Fig. 5. The curves of Fig. 3d are all tangent to the vertical axis at the same point a; for each of these sets of roots the distance Oa is the geometrical mean value of all four roots.

The general case of four complex roots is illustrated by Fig. 4 for the numerical values $\alpha_1 = -1+i10$, $\alpha_2 = -1-i10$, $\alpha_3 = -2+i15$, $\alpha_4 = -2-i15$. For all complex roots the poles must also be complex; the pole with positive imaginary part must lie in the region bounded by the curve $\Gamma = \Gamma_1+\Gamma_2+\ldots+\Gamma_7$. This curve is tangent to the vertical axis at the point a, and tangent to a vertical line at the left at the point d. The largest absolute value of any point in the domain occurs at the point c, and the smallest at f; these two points are the points of tangency of the curve Γ with circles about the origin as center. The curve Γ is tangent at the point e to a circle through the origin having its center on the real axis. The coordinates of these points are all given in Table V.

The general case of four real roots is illustrated by Fig. 5 for the numerical values $\alpha_1 = -1$, $\alpha_2 = -2$, $\alpha_3 = -5$, $\alpha_4 = -7$. The domain of complex poles is bounded by the curve Γ, with the critical points defined and labeled as in Fig. 4. The domain of complex poles is bounded in part by two segments on the real axis, one lying in the interval between α_1 and α_2, the other between α_3 and α_4. Approximately, these segments are from -1.13 to -1.93 and from -5.13 to -6.70, for this numerical example. The points on these segments are in the domain of poles, corresponding to double real poles. The domain of real poles is shown by the graph below the axis, each point of this graph representing two real values, the two points on the real axis reached by following the $\pm 45°$ lines through the point. The domain of real poles is bounded by the continuation of the curve Γ and the tangent lines corresponding to the four roots. This gives two corners associated with the two segments on the real axis, and an isolated rectangle. Corresponding to the points in the rectangle, one pole may be chosen anywhere in the range from α_1 to α_2, and the second pole anywhere in the range from α_3 to α_4. Both poles may be chosen in the range from α_1 to α_2, or in the range from α_3 to α_4, with certain restrictions as shown by the figure, since the curve Γ cuts off the points of the triangles. The two corners and the rectangle are shown by Fig. 5a on a larger scale, with greater accuracy.

In some respects, the case illustrated by Fig. 5 is the most general case, from which all other cases can be obtained by a continuous transformation of the roots. Two of the adjacent real roots may be brought together to a single double root; the corresponding boundary curve then shrinks to a cusp at this point on the real axis, and the rectangle

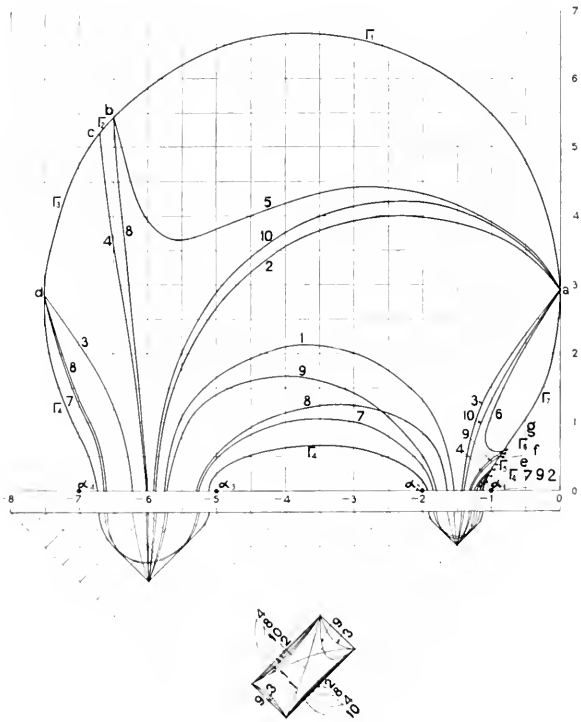


Fig. 5 — Domain of the poles of the driving-point impedance of a two-mesh circuit with four real roots, showing the portions of the domain realizable by each network listed in Table III.

in the auxiliary diagram narrows down to a single line segment. Then if the other two real roots are brought together, the boundary curve has a second cusp and the domain in the auxiliary diagram shrinks to a single isolated point. It now, one of the pairs of equal real roots is separated into a pair of conjugate imaginary roots, the

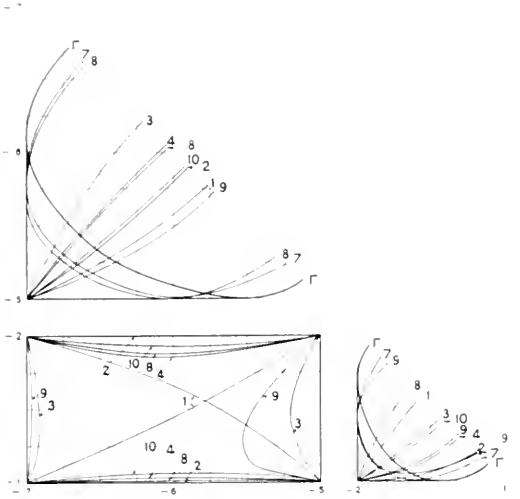


Fig. 5a.—Domain of real poles of Fig. 5, on larger scale.

corresponding cusp is rounded off away from the axis, and the point in the auxiliary diagram vanishes. When the other pair of equal real roots separates into conjugate complex roots, the case illustrated by Fig. 4 is obtained. As one pair of complex roots approaches the imaginary axis, the domain narrows until, for one pair of roots on the vertical axis, the domain shrinks to a circular arc as illustrated by Fig. 3a. This sort of transformation may be followed through in different ways in order to obtain any desired distribution of the roots.

The complete domains are unique; that is, any one domain is given by only one set of roots.

Every domain includes the points corresponding to the roots for which the domain is defined. For these points, that is, for a pole coinciding with a root, the impedance expression has a common factor

in numerator and denominator. When both poles coincide with roots the corresponding impedance expression can be obtained by means of a one-mesh circuit.

5. Two-Mesh Circuits and Associated Networks

The second object of this paper is the determination of the networks realizing any specified driving-point impedance which satisfies the conditions established in the first part of the paper. It is necessary to find the number, character, and arrangement of the elements in these networks, as well as to find the values of these elements.

Thus the problem met in this investigation differs from the usual network problem in that it calls for the determination of the elements of a network which has a certain specified impedance, instead of calling for the determination of the impedance of a network which has certain specified elements.

The most general two-mesh circuit has three branches connected in parallel, each branch containing resistance, capacity, and self-

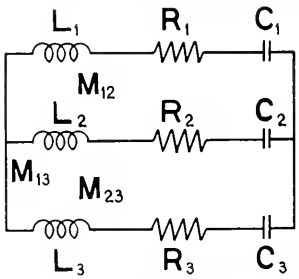


Fig. 6—Most general two-mesh circuit consisting of resistances, capacities, and inductances.

inductance, with mutual inductance between each pair of branches, as shown by Fig. 6.

The most general network under consideration is, therefore, the network obtained by opening one branch of this two-mesh circuit, as shown by Fig. 7. All the networks considered are special cases of this general network, obtained by making a sufficient number of the elements either zero or infinite. If, in particular, all the elements in one branch are replaced by a short circuit, the network splits up into two separate sections connected essentially only by mutual inductance, as shown by Fig. 7a.

It is convenient to limit this investigation to the determination of those networks which, without superfluous elements, realize any driving-point impedance having arbitrarily assigned roots. A network is considered to have superfluous elements if there exist other

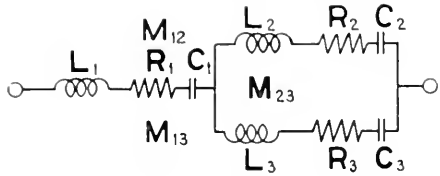


Fig. 7.—Most general network obtained by opening one branch of a two-mesh circuit.

networks with fewer elements which, individually or collectively, realize the same range of possible impedances. Impedances with zero, pure imaginary, or infinite roots can be realized by the limiting cases of these networks.

A network realizing an impedance with arbitrarily assigned roots must consist of at least five elements, one resistance, two capacities,

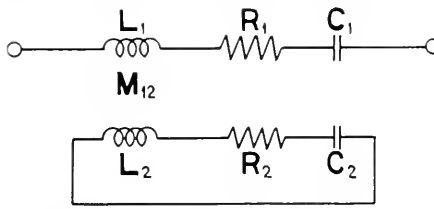


Fig. 7a.—Special case of Fig. 7, obtained by replacing the elements of one branch by a short circuit.

and two self-inductances, in order that the numerator of the impedance expression (1b) may contain odd powers of λ, a constant term, and a term in λ^4, respectively.

Since the general expression for the driving-point impedance contains essentially seven constants which may be assigned arbitrarily, subject to the restrictions already established, it is to be expected that the entire range of possible impedances can be realized by one or more networks consisting of seven elements only. This proves to be the case. Hence all networks with more than seven elements

contain superfluous elements. It is also to be expected that one additional condition must be satisfied by the roots and poles in order that an impedance may be realized by a six-element network, and two additional conditions for a five-element network.

Accordingly, a census has been made of all networks consisting of not more than seven elements, each network containing at least one resistance, two capacities, and two self-inductances. This census is shown by Table III.

Each two-mesh circuit meeting these requirements as to the number of elements is represented in symbolical form in Table III. The letters L, R, and C, printed in the first, second, or third lines of the symbol, indicate the presence of self-inductance, resistance, and capacity in the first, second, or third branches of the circuit, respectively. The letter M is printed in the two lines of the symbol corresponding to the two branches which are connected by a mutual inductance. Thus the first circuit in the table is represented by the symbol

$$\begin{matrix} LRCM \\ L\ \ CM \\ L \end{matrix}$$

which indicates self-inductance, resistance, and capacity in the first branch, self-inductance and capacity in the second branch, and self-inductance in the third branch, with mutual inductance between the first two branches.

Three networks, in general, are obtained from each of these circuits by opening each of the three branches. If two of these branches are alike, only two distinct networks are obtained. If one branch of a circuit is a short-circuit, there being no elements assigned to that branch, the network obtained by opening one of the other branches is of the type shown by Fig. 7a; if the short-circuited branch is opened, the network consists simply of the parallel combination of the other two branches.

With circuits represented in this symbolical manner, there is, opposite each line of the symbol, a reference to the domain of poles indicating the portion of the domain realizable by the network obtained by opening the corresponding branch. Two like branches in a circuit are bracketed together with a single reference mark, since they each give the same network. The entire domain is indicated by Σ; the boundary curve of the domain by Γ, this being divided into seven segments, Γ_1, Γ_2, . . . , Γ_7; ten critical lines in the domain by the

TABLE III

Networks Realizing the Driving-Point Impedance of a Two-Mesh Circuit

1.	*LRCM* *L C M* *L*	(6a) (6c) (7)		13	*LRC* *LRC* *L*	(2) $1-8-\Gamma_2-\Gamma_3$		20.	*LRCM* *L C M*	(a) (c)		25.	*LRC* *L C* *L*	(a) (c) $\Gamma_2, 1$		34	*LRC* *L C*	* *
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
2.	*LRCM* *L C* *L M*	(6a) (6c) (7)		14	*LRC* *LR* *L C*	$3-4-6-\Gamma_4-\Gamma_5$ $3-5-10-\Gamma_2-\Gamma$ $4-7-9-\Gamma-\Gamma_1$		21.	*LRCM* *L M* *C*	(a) Γ (c)		26.	*LR* *L C* *L C*	(a) Γ_3, Γ_5		35	*LRC* *L* *C*	* * **
3.	*LRC* *L CM* *L M*	(6a) (6c) (7)		15.	*LRC* *L C* *L C*	(6a) $(\Gamma_2, \Gamma_3, \Gamma_4, \Gamma_5)$		22.	*LR M* *L CM* *C*	(a) Γ Γ		27.	*LRC* *LRC*	* 1		36	*LR* *L C* *C*	a d g
4	*LR M* *L CM* *L C*	(6a) (7) (7)		16.	*LRC* *LRC* *R*	(4) $1-10$		23.	*L CM* *L CM* *R*	Γ (a)		28.	*LRC* *LR* *C*	3 3 9		37.	*L C* *L C* *R*	c a
5.	*LR* *L CM* *L CM*	(6a) (7)		17.	*LRC* *LR* *RC*	$2-3-7$ $3-5-8$ $2-6-9$		24.	*L CM* *L M* *RC*	(c) Γ (a)		29.	*LRC* *L C* *R*	4 4 10		38	*L C* *L* *RC*	e b a
6	*LRCM* *LRCM*	(2) (2)		18.	*LRC* *LRC* *C*	(3) $1-9-\Gamma-\Gamma_5$						30.	*LRC* *L* *RC*	2 8 2				
7.	*LRCM* *LR M* *C*			19	*LRC* *L C* *RC*	$2-4-5-\Gamma_2-\Gamma$ $4-7-8-\Gamma_3-\Gamma_4$ $2-6-10-\Gamma-\Gamma_1$						31	*LR* *L C* *RC*	5 7 6				
8.	*LRCM* *L CM* *R*											32.	*LRC* *L C* *C*	(a) (d) $\Gamma, 1$				
9	*LRCM* *L M* *RC*	(2) (2)										33.	*L C* *L C* *RC*	$\Gamma, 1$ (a)				
10	*LR M* *L CM* *RC*																	
11	*LRCM* *L CM* *C*	(6a) (7) 1																
12	*L CM* *L CM* *RC*	(7) (a)																

* Equivalent to a single-mesh network.

** Network by which only impedances having one pair of imaginary roots can be realized.

† The impedances realized by this network can be realized by other networks with one less element.

‡ The impedances realized by this network can be realized by other networks with two less elements.

numbers 1, 2, . . . , 10; and seven critical points by the letters a, b, . . . , g, as illustrated by Figs. 4 and 5.

Networks with superfluous elements are indicated by placing parentheses around the corresponding reference mark, single parentheses for one superfluous element and double parentheses for two. In order that a seven-element network may contain no superfluous elements it must give the entire domain or a region in it, a six-element network a critical line, and a five-element network a critical point.

That is, an impedance with arbitrarily assigned roots, and with a pole chosen arbitrarily in the domain corresponding to these assigned roots, can be realized with the minimum number of elements only by a seven-element network. If the pole is chosen so as to satisfy one additional condition, namely, chosen at a point on one of the critical lines of the domain (including the boundary curve), the impedance can be realized by the six-element network giving that line. If the pole is chosen so as to satisfy two additional conditions, namely, chosen at one of the critical points, the impedance can be realized by the corresponding five-element network.

The conditions for the critical lines and for the critical points are given by Tables IV and V, respectively, in terms of the coefficients of the impedance.

TABLE IV

Critical Lines

1. $a_3^2b_1^2+(4a_1a_4-2a_2a_3)b_1b_2-2a_1a_2b_1b_3-(4a_0a_4-a_2^2)b_2^2+(4a_0a_3-2a_1a_2)b_2b_3+a_1^2b_3^2=0.$

1. $(8a_3a_4^2-4a_2a_3a_4+a_3^3)b_1^4-(16a_0a_4^2+2a_1a_3a_4-4a_2^2a_4+a_2a_3^2)b_1^2b_2$
$+(8a_0a_3a_4-4a_1a_2a_4+a_1a_3^2)b_1^2b_3+(8a_0a_3a_4-4a_1a_2a_4+a_1a_3^2)b_1b_2^2$
$-6(a_0a_3^2-a_1^2a_4+a_1a_2a_3)b_1b_2b_3-(8a_0a_1a_4-4a_0a_2a_3+a_1^2a_3)b_1b_3^2$
$+(a_0a_3^2-a_1^2a_4)b_2^3-(8a_0a_3a_4-4a_0a_2a_3+a_1^2a_3)b_2^2b_3+(16a_0^2a_4+2a_0a_1a_3$
$-4a_0a_2^2+a_1^2a_2)b_2b_3^2+(8a_0^2a_3-4a_0a_1a_2+a_1^3)b_3^3=0.$

2. $2a_4b_1b_3-a_3b_1b_2^2-a_4b_2^2+a_3b_2^2b_3-a_2b_2b_3^2+a_1b_3^4=0.$

3. $a_3b_1^4-a_2b_1^3b_2-a_1b_1^2b_3+a_1b_1b_2^2+2a_0b_1b_2b_3-a_0b_2^4=0.$

4. $a_4b_1^3b_2-a_3b_1^2b_2^2+a_2b_1b_2^3-a_1b_2^4-a_4b_1^2b_3+a_3b_1b_2b_3-a_2b_2^2b_3+a_1b_3^2=0.$

5. $a_4^2b_1^4b_2+a_3^2b_1^4b_3+(a_3a_4-a_4^2)b_1^3b_2^2-(a_1a_4-a_2a_3)b_1^2b_2b_3$
$-2a_1a_3b_1^2b_3^2-a_1a_4b_1b_2^3+a_1a_3b_1b_2^2b_3+(a_0a_3-a_1a_2)b_1b_2b_3^2$
$+a_1^2b_1b_3^3+a_1a_4b_2^4-a_1a_3b_2^3b_3+a_1a_2b_2^2b_3^2-a_0a_1b_2b_3^3=0.$

6. $a_4^2b_1^4b_3-a_3^2b_1^3b_2^2+a_3a_4b_1^3b_2^3+a_4a_5b_1^3b_3^2+2a_1a_2b_1^2b_2^2b_3$
$+(a_1a_4b_1b_2^4-a_1a_5b_1b_2^3b_3+(a_0a_4-a_1a_3)b_1b_2^2b_3^2-a_1^2b_1b_3^4-a_0a_4b_2^4$
$+a_2a_5b_2^3b_3-a_2^2b_2^2b_3^2+a_1^2b_2b_3^3-a_0a_5b_2^2b_3^3=0.$

7. $a_4^2b_1^4b_3-a_3^2b_1^3b_2+a_3a_4b_1^3b_2^2+(a_4a_5+a_3a_5)b_1^3b_2b_3-2a_1a_2b_1^2b_2^2b_3^2$
$+a_1a_4b_1b_2^4-a_1a_5b_1b_2^3b_3+(a_0a_4+a_1a_2)b_1b_2^2b_3^2-a_1^2b_1b_3^4-a_0a_4b_2^4$
$+a_2a_5b_2^3b_3-a_0a_5b_2^2b_3^2+a_0a_5b_2b_3^3=0.$

8. $(8a_1a_4^2 - 4a_1a_2a_3a_4 + a_1^2b_1^2 - (8a_1a_3a_4^2 + 2a_1a_2a_4 - 4a_2^2a_4 + a_2a_3^2)b_1^4b_2$
$+ (4a_1a_3^2a_4 + a_1a_2^2a_3 - 4a_1^2a_2a_4)b_1^4b_1 + (2a_1a_3a_4 + a_1^2a_2)b_1^4b_2^2$
$- (4a_2a_3a_4 - 6a_1^2a_4)b_1^3b_2b_1 + (4a_1a_3a_4 - a_1^2a_2)b_1^3b_1^2 - (a_1a_3^2 - 4a_1a_2a_4$
$- a_1^2a_4)b_1^2b_2^3 - (8a_1a_3a_4^2 + a_1^2a_2^2)b_1^2b_2^2b_1 + (2a_1a_3a_4 + a_1^2a_2)b_1^2b_2b_1^2$
$- a_1^3b_1^2b_1^3 - 2a_1a_3a_4^2b_1b_2^4 + (2a_1^2a_4 + 2a_1a_3a_4)b_1b_2^3b_1$
$- (a_1^2a_4 + 2a_1a_3a_4)b_1b_2^2b_1^2 + 2a_1a_4^2b_1b_2b_1^3 + a_1^2a_4b_2^4 - a_1^2a_4b_1^2b_2$
$+ a_1^2a_3b_2^2b_3^2 - a_1^2a_4^2b_2^2b_1^2 = 0.$

9. $a_3a_4^2b_1^4b_2^2 - 2a_1^2a_4b_1^4b_2b_3 + a_1^4b_1^4b_3^2 - a_3a_4^2b_1^4b_2^2 + (a_1a_4^2 + 2a_2a_3a_4)b_1^2b_2^2b_3$
$- (2a_1a_3a_4 + a_2a_3^2)b_1^2b_2b_3^2 - (4a_1a_4a_4 - a_1a_3^2)b_1^3b_1^4 + a_1a_3^2b_1b_2^4$
$- (2a_1a_4^2 + 2a_1a_3a_4 - b_1b_2^2b_3 + (8a_1a_3a_4 + a_1a_2^2)b_1b_2^2b_3^2$
$+ (4a_1a_3a_4 - 6a_1a_3^2)b_1b_2b_3^3 - (4a_1a_3a_4 - 4a_1a_4^2 + a_2^2a_3)b_1b_3^4$
$- a_1a_4^2b_2^2 + 2a_1a_3a_4b_2^3b_3 - (4a_1a_3a_4 + a_1a_2^2 - a_1^2a_4)b_2^3b_1^2$
$- (2a_1a_4^2 - 4a_1a_3a_4 + a_1^2a_3)b_2^2b_1^3 + (8a_1^2a_4 + 2a_1a_3a_4 - 4a_1a_2^2 + a_1^2a_2)b_2b_1^4$
$- 8a_1^2a_3 - 4a_1a_2a_4 + a_2^2)b_1^5 = 0.$

10. $a_3^2a_4b_1^4b_2 + a_3^2b_1^4b_3 - a_3^2b_1^4b_2^2 - 2a_1a_3a_4b_1^3b_2b_3 - 3a_1a_3^2b_1^3b_2^2$
$+ (4a_1a_4^2 - a_2^2a_4 + a_2a_3^2)b_1^2b_2^3 + (4a_1a_3a_4 + 3a_1a_3^2 - a_2^2a_3)b_1^2b_2^2b_3$
$- (a_2a_3^2 - a_2^2a_4)b_1^2b_2b_3^2 + 3a_1^2a_3b_1^2b_2^4$
$- (4a_1a_3a_4 - 2a_1a_4^2 + a_1a_2^2)b_1b_2^3 - (4a_1a_4a_4 + 3a_1^2a_3 - a_1a_2^2)b_1b_2^2b_3^2$
$+ 2a_1a_4^2a_1b_1b_2b_3^2 - a_1^2b_1b_3^4 + (a_2a_4^2 - a_2^2a_4b_2^5 + (4a_1a_4a_4 - 2a_1a_2a_3$
$+ a_1^2a_3)b_2^4b_3 - (4a_1^2a_4 - a_1a_2^2 + a_1^2a_2)b_2^3b_3^2 + a_1^4b_2^2b_3^3 - a_2a_4^2b_2b_3^4 = 0.$

TABLE V

Critical Points

Point	Coordinates	
	$\frac{b_2}{b_1}$	$\frac{b_3}{b_1}$
a	0	$\frac{a_1}{a_4}$
b	$\frac{a_1a_3^2 + a_1^2a_4 - a_1a_2a_4}{a_3(a_1a_4 - a_2a_3)}$	$\frac{a_1a_3a_4 + a_1a_2a_4 - a_2^2a_3}{a_3(a_1a_4 - a_2a_3)}$
c	$\frac{a_1}{2a_0} - \frac{2a_0a_3 - a_1a_2}{2a_0\sqrt{a_2^2 - 4a_0a_4}}$	$\frac{1}{2a_0}(a_2 + \sqrt{a_2^2 - 4a_0a_4})$
d	$\frac{a_1}{a_0}$	$\frac{a_1a_2 - a_0a_3}{a_0a_4}$
e	$\frac{a_2^2}{a_2a_3 - a_1a_4}$	$\frac{a_3a_4}{a_2a_3 - a_1a_4}$
f	$\frac{a_1}{2a_0} + \frac{2a_0a_3 - a_1a_2}{2a_0\sqrt{a_2^2 - 4a_0a_4}}$	$\frac{1}{2a_0}(a_2 - \sqrt{a_2^2 - 4a_0a_4})$
g	$\frac{a_1a_3^2 + a_1^2a_4 - a_1a_2a_3}{a_3a_1a_4 + a_3a_2a_1 - a_1a_2^2}$	$\frac{a_1(a_1a_4 - a_2a_3)}{a_3a_1a_4 + a_3a_2a_1 - a_1a_2^2}$

These critical lines and points are illustrated, for numerical cases, by Figs. 4 and 5. The graph showing the domain of real poles in Fig. 5 is inaccurate to the extent that the critical lines have been spread somewhat apart from each other in order to show the sequence in which they occur. The actual curves are shown accurately drawn and on a larger scale in Fig. 5a. Even on this scale, Curve 2 cannot be distinguished from the side of the rectangle.

The diagrams for the domain of complex poles, as illustrated by Figs. 4 and 5, are approximately symmetrical with respect to the interchanging of inductances and capacities, with corresponding interchanges in all the curves and formulas. Thus b and g correspond, c and f, d and e, 2 and 3, 5 and 6, 8 and 9, α_1 and α_4, α_2 and α_3; while a, 1, 4, 7, and 10 remain unchanged. In the domain of real poles shown by Fig. 5, this symmetry does not appear. The explanation of this apparent discrepancy is as follows: Upon interchanging inductances and capacities, the values of the roots are changed to their reciprocals. Thus Fig. 5 is symmetrical with the corresponding figure drawn for the case of roots equal to -1, -1.2, -1.5, and -1.7, and thus symmetrical with the figure drawn for roots at -7, -7.2, -7.5, and -1, since the relative distribution of the roots is the same. This set of roots differs not very considerably from the original set of roots, in reverse order. In the main, therefore, the two figures may be expected to be approximately the same, that is, the original figure symmetrical with itself. In the rectangle, however, very small numerical changes in the constants make relatively large changes in the curves; so it is not surprising to find a lack of symmetry here. If the roots are assigned so that the product of two roots is equal to the product of the other two, there will be true symmetry in the corresponding diagram.

Table III lists 38 circuits, giving a total of 102 networks. Of these networks, three are essentially the equivalent of networks obtained from a one-mesh circuit, one realizes only those impedances which have one pair of pure imaginary roots, and, of the 98 remaining, 11 have superfluous elements. This leaves a total of 57 networks, of which 11 realize the entire domain as given by Theorem II, 12 realize regions in the domain as given by Theorem III, 23 realize critical lines in the domain, and 11 realize critical points.

The eleven networks of Theorem II are included in the first column of Table III and shown in detail by Fig. 4. Formulas for the computation of their elements are given by Table I. Thus the values of these elements can be computed directly in terms of the coefficients of the impedance expression as stated in the form (11b). The following method of computation is convenient: First compute d as the root

of the quadratic equation (21), which is repeated at the bottom of the table. Then find c by subtracting this value of d from a_2. Next compute I_1, I_2, and I_3, assigning signs so that the identity $b_1I_1+b_2I_2+b_3I_3=0$ is satisfied; this is possible since the equation for d was obtained by rationalizing this relation among the I's. There are, in general, two sets of signs for which this identity is satisfied; it is immaterial which set is chosen since the signs of all the I's may be changed without changing the values of any of the elements. Then compute U_1, U_2, and U_3, assigning positive values to each of these. With the values of all these quantities determined, the values of the elements of the networks can be calculated directly from the formulas given in the body of the table. If this solution turns out to be impossible, that is, if the value of an element is found to be negative or complex or if the value of a mutual inductance is found to be greater than the square root of the product of the associated self-inductances, it means that the conditions upon the roots and poles are not satisfied. If the conditions established in the first part of this paper are satisfied, the solution is possible.

These formulas give all the special cases of the eleven networks automatically, that is, the values of the appropriate elements will turn out to be zero or infinite, as the case may be. Since each of these eleven networks covers the entire domain, they are all mutually equivalent at all frequencies. These are the only networks without superfluous elements which cover the entire domain, that is, any network covering the entire domain must be one of these eleven or a network obtained from one of these by introducing additional elements. Each of the eleven contains just seven elements; thus the prediction that a seven-element network would cover the entire domain is verified. The three remaining networks of this same type, one from Circuit 6 and two from Circuit 9 of Table III give special cases only, in the sense that each of these can realize only those impedances which have a pole lying on Line 2; thus each of these three contains a superfluous element, since all the points on Line 2 can be realized by six-element networks, as shown in the fourth column of the table.

Network 4 of Fig. 1 is of particular interest since it consists simply of two branches in parallel, each containing resistance, capacity, and self-inductance, with mutual inductance between them.[10] By Theorem II, this network can be made equivalent to any network whatsoever obtained from a two-mesh circuit.

[10] It will be shown in a subsequent paper that any driving-point impedance of an n-mesh circuit can be realized by a network of n branches in parallel, each branch containing resistance, capacity, and self-inductance, with mutual inductance between each pair of branches.

The twelve networks of Theorem III are included in the second column of Table III and shown in detail by Fig. 2. Formulas for the computation of their elements are given by Table II. The values of the elements can be computed by the same rule as that given above for Table I.

Each of these twelve networks realizes those impedances which have poles lying in a certain restricted area or region of the entire domain of possibilities, as indicated for each network in the table by a specification of the boundary curves of the area. For each particular impedance in the domain various sets of these twelve networks are mutually equivalent. Some points in the domain cannot be realized by networks without mutual inductance. Of the remaining points, each is realizable, in general, by at least three, and by not more than five, of these twelve networks. This region of the domain which is realizable without mutual inductance is covered, with no overlapping, by each of the four following sets of networks: 13, 17, and 21; 13, 18, and 22; 14, 17, and 23; 15, 19, and 21; the numbers refer to the networks of Fig. 2.

That portion of the domain which cannot be realized by networks without mutual inductance comprises the three regions bounded by F_1 and 5, F_4 and 7, and F_7 and 6, respectively, as illustrated by Figs. 4 and 5.

The third and fourth columns of Table III show a total of 23 networks, each with six elements, realizing lines in the domain. Of these, eleven are derived as special cases of the networks of both Figs. 1 and 2, six as special cases of Fig. 1 but not of Fig. 2, and six as special cases of Fig. 2 alone. The fifth column of the table shows the eleven networks, each with five elements, realizing points in the domain.

6. Formulas for Calculation of General Network

Formulas for the calculation of the values of the elements of the general network of Fig. 7 are given in Theorem IV. These are given in the form of nine equations (7) (15), inclusive, involving the twelve elements of the network and two parameters, d and k. The parameter d, however, is fixed by the impedance, since the left-hand members of equations (13) (15) satisfy the identity (20). Upon substituting the right-hand members in the identity and rationalizing, equation (21) is obtained, this being a quadratic equation in d with coefficients which are functions of the known coefficients of the impedance. Since d is fixed in this way, there are essentially eight equations in thirteen variables, the twelve elements and the arbi-

trary parameter k. In general, therefore, five of the elements may be specified, or five relations among the elements; whereupon the equations can be solved. Thus it is to be expected that a seven-element network will realize, in general, any specified driving-point impedance.

This method of solution is best illustrated by considering a particular case. Take, for example, the derivation of the formulas for Network 1 of Fig. 1, as given by Table I. This is the special case of the general network of Fig. 7 obtained by making $L_1 = R_1 = C_1^{-1} = M_{12} = M_{13} = 0$. Substituting these values, together with the notation of Table I, equations (7)–(15) become

$$\begin{aligned}
L_2L_3 - M_{23}^2 &= a_0k^2,\\
R_2R_3 &= dk^2,\\
D_2D_3 &= a_4k^2,\\
L_2 + L_3 - 2M_{23} &= b_1k^2,\\
R_2 + R_3 &= b_2k^2,\\
D_2 + D_3 &= b_3k^2,\\
R_2D_3 - R_3D_2 &= T_1k^4,\\
D_2L_3 - D_3L_2 - (D_2 - D_3)M_{23} &= T_2k^4,\\
L_2R_3 - L_3R_2 - (R_3 - R_2)M_{23} &= T_3k^3.
\end{aligned}$$

Eliminating R_2, R_3, D_2, and D_3 from the second, third, fifth, sixth, and seventh of these equations, the value of k is found to be equal to $\pm U$, T_3. Knowing the value of k, the equations may then be solved for the seven elements, obtaining the results given in Table I. The two sign choices for k in this example correspond to the possibility of interchanging branches 2 and 3 in the network. The values given in Table I are computed for k taken with the negative sign.

In the general solution, the parameter d is obtained from the quadratic equation (21). The explicit solution of this equation is

$$d = \frac{2a_4b_1^2 + a_2b_2^2 + 2a_0b_3^2 - a_3b_1b_2 - 2a_2b_1b_3 - a_1b_2b_3 \pm 2\Delta}{b_2^2 - 4b_1b_3}, \tag{34}$$

where

$$\begin{aligned}
\Delta^2 &= a_4^2b_1^4 + a_4a_1b_2^4 + a_0^2b_3^4 - a_3a_4b_1^3b_2 - (2a_2a_4 - a_3^2)b_1^3b_3 - a_3a_1b_1b_2^3\\
&\quad - a_4a_3b_2^3b_3 - (2a_0a_2 - a_1^2)b_1b_3^3 - a_0a_1b_2b_3^3\\
&\quad + a_2a_4b_1^2b_2^2 + (a_2^2 + 2a_0a_4 - 2a_1a_3)b_1^2b_3^2 + a_0a_2b_2^2b_3^2\\
&\quad + (3a_1a_4 - a_2a_3)b_1^2b_2b_3 - (4a_0a_4 - a_1a_3)b_1b_2^2b_3\\
&\quad - (3a_0a_3 - a_1a_2)b_1b_2b_3^2, \qquad (35)\\
&= a_0^2(\alpha_1^2b_1 + \alpha_1b_2 + b_3)(\alpha_2^2b_1 + \alpha_2b_2 + b_3)\\
&\quad (\alpha_3^2b_1 + \alpha_3b_2 + b_3)(\alpha_4^2b_1 + \alpha_4b_2 + b_3), \qquad (36)\\
&= a_0^2b_1^4(\alpha_1 - \beta_2)(\alpha_1 - \beta_3)(\alpha_2 - \beta_2)(\alpha_2 - \beta_3)\\
&\quad (\alpha_3 - \beta_2)(\alpha_3 - \beta_3)(\alpha_4 - \beta_2)(\alpha_4 - \beta_3). \qquad (37)
\end{aligned}$$

In the case of real and distinct poles, formula (31) gives, in general, two positive values of d satisfying the necessary conditions (4)–(6), and thus two solutions for any particular network. For complex poles, only one such value of d is obtained, and there is thus a unique solution in each case. For real and equal poles, $b_2^2-4b_1b_3=0$, and so formula (34) does not apply directly; in this case, however, (21) reduces to a linear equation in d, so that the solution can be readily found.

An obvious necessary condition for a solution is that $\Delta^2\geq 0$, for otherwise the value of d would be complex. This condition is satisfied for any choice of poles provided there is not an odd number of real roots lying between two real poles. Thus for the case of all complex roots or for the case of complex poles with any choice of roots this condition is automatically satisfied. It is interesting to note that an impedance expression with poles failing to satisfy this condition cannot be realized by any network with positive or negative resistances, capacities, and inductances; it can be realized only by a network with elements having complex values.

7. Networks with Negative Resistances

If negative resistances are allowed in the two-mesh circuit, the only change necessary in the statement of the results of this investigation, as given in Theorems I–IV, is the removal of the restrictions $\alpha_1+\alpha_2\leq 0$, $\alpha_3+\alpha_4\leq 0$, $\beta_2+\beta_3\leq 0$, and $d\geq 0$. This removes the restriction of the real part of each root and pole to negative or zero values. The removal of the restriction on d adds to the domain of poles, considered in the x, y plane, all the ellipses of the family $-\infty<d<0$, thus filling out the region above the critical parabola (33), together with the corners in the case of real roots. In the u, v plane the domain comprises the entire upper half of the complex plane and, in the auxiliary diagram, the complete triangular corners and the rectangle, with the provision that the rectangle is not included in the case of two roots positive and two negative.

By means of a two-mesh circuit employing negative resistances, any impedance expression of the form (1) can be realized, with roots arbitrarily assigned in conjugate pairs or in real pairs, subject only to the condition that the number of positive roots is even, and with any pair of complex poles or with a pair of real poles lying anywhere in the ranges from the first to the second real roots and from the third to the fourth real roots, arranged in order of magnitude, subject only to the condition that both poles must be positive or both negative.

The network diagrams and all the formulas for the calculation of the elements remain unchanged.

8. Mathematical Proof

The circuits treated in this investigation are special cases of the general circuit which has any number of terminals m connected in pairs by $m(m-1)/2$ branches, each of which consists of a self-inductance, a resistance, and a capacity in series, with mutual inductance between each pair of branches. The only restrictions imposed are those inherent in all electrical circuits, namely, that the magnetic energy, the dissipation, and the electric energy are each positive for any possible distribution of currents in the branches. Circuits with any arrangement of elements in series or in parallel or in separated meshes can be derived as limiting cases of this general circuit by making a sufficient number of the inductances, resistances, and capacities either zero or infinite.

This general circuit connecting m terminals or branch-points has $n=(m-1)(m-2)/2$ degrees of freedom, that is, n independent meshes. The discriminant[11] of the circuit is the determinant Δ having the element Z_{jk} in the jth row and kth column, Z_{jk} being the mutual impedance between meshes j and k (self-impedance when $j=k$), the determinant including n independent meshes of the circuit.

The driving-point impedance in the qth mesh S_q is equal to the ratio Δ/Δ_{qq}, where Δ_{qq} is the cofactor of the element in the qth row and qth column of the determinant Δ. In general, the cofactor of the product of the elements located at the intersection of rows $j, q, s, \ldots$ with columns $k, r, t, \ldots$, respectively, will be denoted by $\Delta_{jk\,qr\,st\,\ldots}$.

The determinant Δ for the general circuit described above is of order n with the element

$$Z_{jk}=iL_{jk}p+R_{jk}+(iC_{jk}p)^{-1} \tag{38}$$

where L_{jk}, R_{jk}, and C_{jk} are the inductance, the resistance, and the capacity, respectively, common to the two meshes j and k. The inductance L_{jk} includes, therefore, the self-inductances of the branches common to the two meshes together with the mutual inductances connecting each branch of one mesh with each branch of the other mesh. The determinant is symmetrical, that is $Z_{jk}=Z_{kj}$, since $L_{jk}=L_{kj}$, $R_{jk}=R_{kj}$, and $C_{jk}=C_{kj}$.

[11] A complete discussion of the solution of circuits by means of determinants has been given by G. A. Campbell, *loc. cit.*, pages 883–886.

These coefficients L_{jk}, R_{jk}, and C_{jk} are subject to the energy conditions stated above, namely, that the magnetic energy, the dissipation, and the electric energy,

$$\frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} L_{jk}i_j i_k,\quad \sum_{j=1}^{n}\sum_{k=1}^{n} R_{jk}i_j i_k,\ \text{and}\ \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n}\frac{1}{C_{jk}}\int i_j dt\int i_k dt, \qquad (39)$$

respectively, are each positive for any possible distribution of the currents (i_j, i_k, . . .) in the branches of the circuits.[12] In other words, the coefficients L_{jk}, R_{jk}, and $1/C_{jk}$ are subject to the condition that the three quadratic forms of which these are the coefficients must be positive for all real values of the variables. All the principal minors of the determinants

$$\begin{vmatrix} L_{11} & L_{12} & \dots & L_{1n} \\ L_{21} & L_{22} & \dots & L_{2n} \\ \dots & \dots & \dots & \dots \\ L_{n1} & L_{n2} & \dots & L_{nn} \end{vmatrix},\quad \begin{vmatrix} R_{11} & R_{12} & \dots & R_{1n} \\ R_{21} & R_{22} & \dots & R_{2n} \\ \dots & \dots & \dots & \dots \\ R_{n1} & R_{n2} & \dots & R_{nn} \end{vmatrix},\ \text{and}\ \begin{vmatrix} \frac{1}{C_{11}} & \frac{1}{C_{12}} & \dots & \frac{1}{C_{1n}} \\ \frac{1}{C_{21}} & \frac{1}{C_{22}} & \dots & \frac{1}{C_{2n}} \\ \dots & \dots & \dots & \dots \\ \frac{1}{C_{n1}} & \frac{1}{C_{n2}} & \dots & \frac{1}{C_{nn}} \end{vmatrix} \qquad (40)$$

are positive or zero by virtue of this condition.[13] This same condition holds for the inductances if the coefficients L_{jk} apply to branches instead of meshes.

By expanding the determinants in the numerator and denominator of the expression for the driving-point impedance given above, we find

$$S_q = \frac{\Delta}{\Delta_{qq}} = \frac{a_0(ip)^n + a_1(ip)^{n-1} + a_2(ip)^{n-2} + \dots + a_{2n-1}(ip)^{-n+1} + a_{2n}(ip)^{-n}}{b_1(ip)^{n-1} + b_2(ip)^{n-2} + \dots + b_{2n-1}(ip)^{-n+1}} \qquad (41)$$

[12] For a recent statement of the energy conditions in this form see L. Bouthillon, *Revue Générale de l'Electricité*, 11, 1922, pages 656–661.

[13] A necessary and sufficient condition that the real quadratic form in n variables

$$\sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}x_j x_k \qquad (a_{jk} = a_{kj}),$$

be positive for all real values of the variables is that each of the n determinants,

$$a_{11},\ \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix},\ \dots,\ \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix},$$

be positive. For a proof of this see, for example, H. Hancock, "Theory of Maxima and Minima," 1917, pages 82–91.

Upon substituting $\lambda = ip$, multiplying numerator and denominator by λ^n, and dropping the subscript q, formula (11) becomes

$$S = \frac{a_0\lambda^{2n} + a_1\lambda^{2n-1} + a_2\lambda^{2n-2} + \cdots + a_{2n-1}\lambda + a_{2n}}{b_1\lambda^{2n-1} + b_2\lambda^{2n-2} + \cdots + b_{2n-1}\lambda} \qquad (12)$$

which may be taken as the most general form of a driving-point impedance. This formula, therefore, gives the impedance of the circuit for any electrical oscillations of the form $e^{\lambda t}$, where λ may have any value, real or complex. Formula (12) may be written in the alternative form

$$S = H\frac{(\lambda-\alpha_1)(\lambda-\alpha_2)(\lambda-\alpha_3)\cdots(\lambda-\alpha_{2n-1})(\lambda-\alpha_{2n})}{\lambda(\lambda-\beta_2)(\lambda-\beta_3)\cdots(\lambda-\beta_{2n-1})}. \qquad (13)$$

Thus there are $2n$ roots of S, regarded as a function of λ, which are the $2n$ resonant points of the circuit. There are also $2n$ poles of S, which are the $2n$ anti-resonant points of the circuit, namely, zero, infinity, and the $2n-2$ resonant points of the circuit obtained by opening the branch in which the driving-point impedance is measured.

Upon setting $n=2$ in equations (13) and (12), formulas (1a) and (1b) are obtained, respectively.

From the fact that the coefficients L_{ik}, R_{ik}, and $1/C_{ik}$ satisfy the quadratic form conditions (39), it can be shown mathematically that the coefficients $a_0, a_1, \ldots, a_{2n}$ of (12) are all positive and that the roots $\alpha_1, \alpha_2, \ldots, \alpha_{2n}$ of (13) have negative real parts.[11] This can also be shown from the fact that the free oscillations of the circuit are of the forms $e^{\alpha_1 t}, e^{\alpha_2 t}, \ldots, e^{\alpha_{2n} t}$. Thus the roots occur in pairs each of which has negative real values or conjugate complex values with negative real parts.

The same restrictions hold for the coefficients $b_1, b_2, \ldots, b_{2n-1}$ and the poles $\beta_2, \beta_3, \ldots, \beta_{2n-1}$ since the denominator of S, with the exception of the factor λ^n, is also the discriminant of a circuit. Thus the general restrictions (2) are obtained.

In order to obtain the necessary and sufficient conditions that a function of the type (1b) represent a driving-point impedance realizable by a two-mesh circuit, set this function equal to the impedance of the most general two-mesh circuit and investigate the conditions which must hold upon the coefficients in order that the two forms may be equivalent.

[11] The mathematical work is identical with the mathematics of the corresponding dynamical problem. A detailed proof is given by A. G. Webster, *loc. cit.*

The discriminant of the most general two-mesh circuit is of the form

$$\Delta = \begin{vmatrix} L_{11}\lambda + R_{11} + D_{11}\lambda^{-1}, & L_{12}\lambda + R_{12} + D_{12}\lambda^{-1} \\ L_{12}\lambda + R_{12} + D_{12}\lambda^{-1}, & L_{22}\lambda + R_{22} + D_{22}\lambda^{-1} \end{vmatrix}, \tag{44}$$

where the three sets of coefficients, using D_{jk} instead of $\frac{1}{C_{jk}}$, are subject to the restriction that the three determinants

$$\begin{vmatrix} L_{11} & L_{12} \\ L_{12} & L_{22} \end{vmatrix}, \quad \begin{vmatrix} R_{11} & R_{12} \\ R_{12} & R_{22} \end{vmatrix}, \text{ and } \begin{vmatrix} D_{11} & D_{12} \\ D_{12} & D_{22} \end{vmatrix} \tag{45}$$

are all positive or zero, as well as L_{11}, R_{11}, and D_{11}. This condition requires L_{22}, R_{22}, and D_{22} also to be positive or zero.

The most general driving-point impedance of a two-mesh circuit may be taken as the impedance in the first mesh of the circuit defined by the discriminant (44). Set $\frac{\Delta}{\Delta_{11}}$ equal to the value of S given by (4b). Expanding into polynomials in λ, and equating coefficients of the numerators and denominators of the two expressions, the following relations are obtained:

$$L_{11}L_{22} - L_{12}^2 = a_0k^2, \tag{46}$$

$$L_{11}R_{22} + L_{22}R_{11} - 2L_{12}R_{12} = a_1k^2, \tag{47}$$

$$L_{11}D_{22} + L_{22}D_{11} + R_{11}R_{22} - 2L_{12}D_{12} - R_{12}^2 = a_2k^2, \tag{48}$$

$$R_{11}D_{22} + R_{22}D_{11} - 2R_{12}D_{12} = a_3k^2, \tag{49}$$

$$D_{11}D_{22} - D_{12}^2 = a_4k^2, \tag{50}$$

$$L_{22} = b_1k^2, \tag{51}$$

$$R_{22} = b_2k^2, \tag{52}$$

$$D_{22} = b_3k^2, \tag{53}$$

where k has any real value other than zero. Introduce the notation

$$R_{11}R_{22} - R_{12}^2 = dk^2, \tag{54}$$

where d is positive or zero. Then, using (46), (54), and (50), eliminate L_{11}, R_{11}, and D_{11} from equations (47), (49), obtaining

$$(L_{12}R_{12} - L_{22}R_{11})^2 = k^2(-dL_{12}^2 + a_1L_{22}R_{22} - a_0R_{22}^2), \tag{55}$$

$$(D_{12}L_{22} - D_{22}L_{12})^2 = k^2[-a_0D_{22}^2 + (a_2 - d)D_{22}L_{22} - a_1L_{22}^2], \tag{56}$$

$$(R_{12}D_{22} - R_{22}D_{12})^2 = k^2(-a_1R_{22}^2 + a_3R_{22}D_{22} - dD_{22}^2). \tag{57}$$

Using (51) (53), eliminate L_{22}, R_{22}, and D_{22} from the right-hand members of (55) (57); extract the square root; rearrange the order of the equations, obtaining

$$R_{11}D_{22} - R_{12}D_{12} = \pm k^2(-a_1b_2^2 + a_3b_2b_3 - db_3^2)^{1/2}, \tag{58}$$

$$D_{12}L_{22} - D_{22}L_{12} = \pm k^2[-a_0b_3^2 + (a_2 - d)b_3b_1 - a_1b_1^2]^{1/2}, \tag{59}$$

$$L_{12}R_{12} - L_{22}R_{11} = \pm k^2(-db_1^2 + a_1b_1b_2 - a_0b_2^2)^{1/2}. \tag{60}$$

Thus conditions (4)–(6) are obtained directly from (58)–(60). The left-hand members of (58)–(60) satisfy the identity

$$\begin{aligned}(R_{12}D_{22} - R_{22}D_{12})L_{22} + (D_{12}L_{22} - D_{22}L_{12})R_{22} \\ + (L_{12}R_{22} - L_{22}R_{12})D_{22} = 0.\end{aligned} \tag{61}$$

Substituting (51)–(53) and (58)–(60) in this identity (61), and rationalizing, equation (3) and its equivalent (24) are obtained.

For the general network of Fig. 7,

$$\begin{aligned} L_{11} &= L_1' + L_2', & L_{12} &= L_2', & L_{22} &= L_2' + L_3', \\ R_{11} &= R_1 + R_2, & R_{12} &= R_2, & R_{22} &= R_2 + R_3, \\ D_{11} &= D_1 + D_2, & D_{12} &= D_2, & D_{22} &= D_2 + D_3, \end{aligned} \tag{62}$$

where L_1', L_2', and L_3' are defined by (17)–(19). For this set of constants, branch 2 is made the branch common to the two meshes; the choice of branch 3 as the common branch would not affect the final formulas. Substituting these values (62) in (46), (51), (50)–(53), and (58)–(60), equations (7)–(15) are obtained directly.

Thus Theorems I and IV are completely proved. Theorems II and III are verified by the actual formulas for the elements given in Tables I and II, and by the census of networks presented in Table III.

I am indebted to Dr. George A. Campbell for inspiring the writing of this paper and for specific advice upon many points, and to Miss Frances Thorndike for the preparation of the tables and figures.

Contributors to this Issue

HALSEY A. FREDERICK, B.S., Princeton University, 1910; E.E., Princeton University, 1912; Engineering Department Western Electric Company, 1912—. Mr. Frederick has been engaged in research and development problems connected with telephone receivers, transmitters, and allied subjects in the investigation of the transmission and reproduction of speech and music.

H. F. DODGE, S.B., Mass. Inst. Tech., 1916; Instructor in electrical engineering, Mass. Inst. Tech., 1916–1917; A.M. in Physics and Mathematics, Columbia University, 1922; Engineering Department of Western Electric Company, 1917—. Mr. Dodge has been engaged in development studies on telephone transmitters.

GEORGE A. CAMPBELL, B.S., Massachusetts Institute of Technology, 1891; A.B., Harvard, 1892; Ph.D., 1901; Gottingen, Vienna and Paris, 1893–96; Mechanical Department, American Bell Telephone Company, 1897; Engineering Department, American Telephone and Telegraph Company, 1903–19; Department of Development and Research, 1919—; Research Engineer, 1908—. Dr. Campbell has published papers on loading and the theory of electric circuits, including electric wave-filters, and is also well known to telephone engineers for his contributions to repeater and substation circuits.

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.

OTTO J. ZOBEL, A.B., Ripon College, 1909; A.M., Wisconsin, 1910; Ph.D., 1911; instructor in physics, 1910–15; instructor in physics, Minnesota, 1915–16; Engineering Department, American Telephone and Telegraph Company, 1916–19; Department of Development and Research, 1919—. Mr. Zobel has made important contributions to circuit theory in branches other than the subject of wave-filters.

KARL K. DARROW, S.B., University of Chicago, 1911, University of Paris, 1911–12, University of Berlin, 1912, Ph.D., in physics and mathematics, University of Chicago, 1917, Engineering Department, Western Electric Company, 1917– . At the Western Electric, Mr. Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

RONALD M. FOSTER, S.B., Harvard, 1917; American Telephone and Telegraph Company, Engineering Department, 1917–19; Department of Development and Research, 1919– .

Index to Volume III

O

P

R

S

T

V

W

Z