THE BELL SYSTEM TECHNICAL JOURNAL

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

VOLUME IV
1925

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

The Bell System Technical Journal

January, 1925

Engineering Cost Studies[1]

By F. L. RHODES

Introduction

THE subject assigned to me in the "Notes Regarding the Program of the Conference" is "The Theoretical Principles of Economic Studies and Their Possible Application in Undergraduate Courses." With your permission, I shall digress somewhat from a literal consideration of this title. I shall not undertake to derive formulae, to set up equations and to obtain maxima and minima from them. The mathematics can readily be obtained from available sources. On the other hand, I shall attempt to outline the field for economic studies in engineering work, using illustrations drawn from telephone engineering practice.

What is an engineering cost study? When you or I reach a decision to purchase a certain pair of shoes, making a selection from an assortment ranging in price from (say) $5 to $15, we have performed, consciously or unconsciously, some of the reasoning of an engineering cost study. Among the factors influencing our decision will be the probable length of service life of different pairs, as well as the ability to extend this by an expenditure, to be made at some future time, for maintenance as represented by new soles and heels, which, perhaps, can be applied economically to a moderately costly pair but not so to the cheapest.

These two elements, depreciation and current maintenance, are factors entering into engineering cost studies but they are not all of the factors. Whether we have the necessary capital in hand, or are obliged to hire or otherwise raise it, the annual cost of the capital must be taken into consideration, and treatment of the matter of depreciation is incomplete without consideration of salvage value and cost of removal.

Thus, unless we pursue our investigation into details that are not ordinarily considered when buying shoes, it is evident that our

[1] Notes of a Talk given at the Bell System Educational Conference, August, 1924.

homely illustration, while serving to center our attention on certain important subjects to be taken up in this paper, falls short in respect of others that can not be neglected in engineering cost studies. Broadly speaking, engineering cost studies deal with the comparative annual costs of alternative projects. Frequently they also involve comparisons of expenditures to be made at different times in the future. They are of value to industrial executives in assisting them to arrive at decisions where several courses of action are open, but they are not the sole guides in arriving at decisions. No hard and fast formulae can take the place of judgment based on experience. Formulae of this nature are properly used as guides to assist judgment.

The necessity for guidance from studies of this kind arises most frequently in a growing plant. The telephone plant always has been, and so far as we can anticipate, will continue to be a rapidly growing thing.

This means that whenever an addition is to be made, the question arises, how much capacity for growth is it most economical to provide for? As an illustration of this, consider with me the problem that arises when it becomes necessary to place somewhere an underground cable. Obviously it would be uneconomical to construct an underground conduit of one duct for this cable and next year or the year after to dig up the street and lay another duct for a second cable and so on in piecemeal, hand-to-mouth fashion.

On the other hand, it would not be economical to estimate the number of cables that would be required in a hundred years, even if we could foresee the needs so far ahead with any degree of certainty, and to place at the outset sufficient ducts to care for all the cables required along that route in the next century, for in that event, the carrying charges on the idle ducts would prove much more expensive, in the long run, than would additions made at infrequent intermediate times. Somewhere between one year and one hundred years is the most economical period for which to provide duct capacity in advance. The determination of this period, based on suitable construction costs, the expected rate of growth in cable requirements, and other factors is one of the useful results obtained from an engineering cost study.

Under our organization, practically all types of plant and equipment are developed by the Central Staff. These are standardized in a range of sizes sufficient to meet all the needs of the business.

The choice of standards and sizes to meet specific situations arising in the field is made by the proper officials of the associated operating companies.

If a piece of apparatus or equipment, correctly designed within itself, is installed in the wrong place, or if a wrong size is selected, loss will result.

Questions of where to place plant and what size to employ, and when to replace existing plant constantly confront the operating engineers in the field. In the telephone business every major construction project is described in what we term an "estimate" which is nothing more or less than a detailed design for the project, embodied in drawings and specifications, accompanied by a carefully prepared estimate of its cost. These estimates originate in the Plant Departments of the Associated Companies and are really the bids of the construction forces for performing the work. These estimates pass through the hands of the Chief Engineer of the Associated Company for his scrutiny and approval before they proceed to the higher officials of that company for final authorization. The Chief Engineer considers these estimates in their relation to the general plans of the Company with reference to the growth of the business and the plant. For many years the chief of the Department of which I am a member, Vice President General John J. Carty, occupied the post of Chief Engineer of the New York Telephone Company, the largest associated company of the Bell System. I have heard him say that when, while occupying that position, an estimate for some specific piece of work came before him for review, he asked himself three questions regarding it:

1. Why do it at all?
2. Why do it now?
3. Why do it this way?

Rigorous proof sufficient to answer these three questions will justify the endorsement of any engineering project, and, furthermore, each question generally involves an engineering cost study.

Fundamental Plans

Of all the engineering cost studies that are made in connection with the telephone industry, none is more far-reaching in its effect than those involved in what we term our "fundamental plans." In order to give a fair idea of the importance of the work done under our fundamental plans, it will be necessary to describe briefly what a fundamental plan is.

In completed form a fundamental plan shows what the general lay-out of the telephone plant in a city is expected to be at some definite time, usually from 15 to 20 years in the future. It shows:

(a) The number of central office districts that will be required to provide the telephone service most economically, and the boundaries of these central office districts.

(b) The number of subscribers' lines to be served by each central office.

(c) The proper location for the central office in each district to enable the service to be given most economically with regard to costs of cable plant, land, buildings and other factors.

(d) The proper streets and alleys in which to build underground conduits in order to result in a comprehensive, consistent and economical distributing system reaching every city block to be served by underground cable.

(e) The most economical number of ducts to provide in each conduit run as it is built.

These are all very definite problems that confront the executives of our Associated Companies when plant extensions are required. Our experience has shown that our fundamental plans reduce guessing to a minimum by utilizing the experience of years in studying questions of telephone growth in order to make careful forecasts on the best possible engineering basis. A few words as to how fundamental plans are made may not be out of place.

The basis of the fundamental plan is what we term a commercial survey, which is a forecast of the future community showing the probable amount, distribution and character of the population and the probable market for various classes of telephone service.

Before making this forecast, it is important to know what are the present conditions as to population and use of the telephone service. To ascertain these facts a census of the community from a telephone point of view is made. Present telephone users are classified into:

Residence Telephones.

Business Telephones in Residence Areas.

Telephones in Business Section.

In analyzing Residence telephones all families are divided among those occupying:

(a) Private Residences.

(b) Two-family Houses.

(c) Apartments.
(d) Lodging Houses.

In each class, subdivisions are made according to the rent paid as it has been found that a close relation exists between rent and the class of telephone service used. Business telephones are divided into 20 or 30 different classes. An important factor in the forecast is the future population of the city, both as a whole and by sections.

This involves, in each particular problem, not only study of the past growth of the city in question, but also careful and detailed comparisons with the growth history of other cities where conditions have been such that the experience in those places is useful in making the prediction for the city being studied.

Having arrived at forecasts, for certain future dates, as to the number of telephone users to be provided for, where they will be located, what character of service they will require, what time of day they will call, and how frequently, and where they will call, it becomes a definite, although intricate engineering problem to determine the most economical number, size and location of buildings and switchboards and the location and size of conduit runs. All of the promising combinations of future offices and districts as indicated by experience and the geographical characteristics of the city, are laid out on working maps and the annual costs are figured. The arrangement which gives the lowest equated annual costs over the period of time for which the study is made is, in general, the one which is adopted. Fundamental plans are reviewed every few years, particularly when some major plant addition, for example, the opening of a new central office, comes up for consideration. In this way we are constantly looking ahead and following a coordinated plan; but this plan is not a rigid, fixed thing. It is modified as frequently as may be necessary to meet the constantly changing requirements. In work of this kind, future expenditures must be given greater or less weight accordingly as they are required to be made in the near future or at some more distant time. This is taken into account by equating future expenditures in terms of their present worth; that is, the sum in hand, at the present time, which, at compound interest, will be just sufficient to provide for the future expenditures when they are required.

Transmission Standards and Studies

An interesting and typical annual cost problem which arises in connection with fundamental plans is that of obtaining a proper cost balance between the circuits employed for subscribers' loops and those employed in interoffice trunk lines. The larger the wire, the better will be the talk. But it will also be more expensive. The first step in solving this problem is to decide how good the transmission must be to afford satisfactory service to the telephone using public. Our present standards are a matter of growth; the accumulated results of long and extensive experience. They are live, working standards constantly being intelligently scrutinized and, when necessary, modified. A discussion of the values of the standards employed would unduly prolong this paper. Therefore, let it suffice, at this time, to state that the telephone offices in a large city, including its environs, may be divided into metropolitan offices and suburban offices; that is, the central business offices separated from the suburban residential offices. Between subscribers in different districts suitable standards of transmission are decided upon.

Before describing this study further, reference must be made to the practical necessity for the standardization of construction materials. Subscribers' loops run in length from a few hundred feet to 3, 4 or 5 miles. If we tried theoretically to make all talks exactly equal in loudness, we should have as many different sizes of wire in our cables as there are different lengths of loop. To reduce the complexity, our cable conductors are of certain standard sizes, which experience has shown are sufficiently close together to meet the needs of the business. These standard sizes, in American Wire Gauge, are Nos. 24, 22, 19, 16, 13 and 10; the three latter not being used in subscribers' loops.

Having adopted standards of transmission and standards of cable conductor sizes, our problem is to obtain the standards of transmission with the standards of cable conductors in the most economical manner.

The method of doing this, in brief, is to figure out the annual costs which would be incurred in doing it a number of different ways and to select the way that gives the lowest annual cost. In this kind of a study, which we call a "loop and trunk" study, it has been convenient to designate the subscribers' loops by their maximum circuit resistance. Adopting this form of designation, it may be assumed, first, that all of the subscribers' loops will have an average transmitting and receiving efficiency as good or better than a 350-ohm loop; as a second assumption, that they will be as good or better than a 400-ohm loop; and, as third and fourth assumptions, 450 and 500-ohm loops, respectively. In assuming, for example, a 350-ohm loop in

No. 24-gauge cable, it is, of course, necessary that all subscribers having loops longer than the amount of No. 24-gauge cable represented by this resistance shall be put in No. 22-gauge or No. 19-gauge cable as may be required.

The transmission losses, both transmitting and receiving, are then computed for the assumed loops. The transmission losses in central office apparatus are constant and known. Subtracting the losses in the offices and in the substation loops for each assumed grade of loop from the transmission standards, leaves the amount of transmission loss which can be allowed in the interoffice trunks corresponding to each limiting grade of subscriber's loop. On the basis of this allowable transmission loss in the trunks and knowing the distances between central offices, we are enabled to fix the size of conductor required in the trunks.

Knowing the grade of loops and trunks required for each of the above assumptions, we can then compute the total annual charge of giving service according to that assumption. If the assumptions have been wisely chosen it will usually work out that the first assumption, that is, a very high grade of subscriber's loop, will not be as economical as some others, due to the relatively high cost of the subscribers' loops taken as a whole. Neither will the last assumption, that is, a very low grade of subscriber's loop, be the most economical, on account of the relatively high cost of the trunks. Somewhere between, however, there will be some assumption which will show the smallest total annual charge.

To find more precisely the most economical arrangement, the various values are plotted with the assumptions as to subscribers' loops forming one set of ordinates and the total annual cost forming the other. The point on the curve representing the lowest annual cost then indicates the proper grade of subscribers' loops to employ. In the case of the longer interoffice trunks, loading is, of course, employed. In the design of toll lines and toll switching trunks generally similar cost balancing methods are employed.

In many cases, the problem can be solved by the determination of what we term "the warranted annual charge" of transmission which may be defined as the annual cost of improving the talking efficiency of the circuit in the cheapest way by a definite small amount. By means of studies of this kind, we obtain a plant closely approximating a balanced cost condition. That is, in such a plant, a dollar can be spent in improving transmission efficiency, no more effectively in one part than in another.

Other Applications of Engineering Cost Studies

From what has already been said, it should not be inferred that the sole application of engineering cost studies is in connection with the problems arising in the operating field. The question whether or not a more efficient piece of apparatus at a higher cost is warranted enters into most of our development problems. The economics of the case lie at the root of our development work in all portions of the plant.

At this point I should like to call attention to the fact that our development work covers not only what are termed "transmission" matters, but also very important problems in switchboards, outside plant and other phases of the business.

The service which we provide is a *communication* service, which involves important problems affecting the means for connecting and disconnecting the parties as well as those other important problems, to which your attention has been particularly directed, relating to the loudness and quality of the transmitted speech.

In cable design, particularly in the case of intercity cables and interoffice trunk cables, the average separation between wires in the cable affects the electrostatic capacity of the circuits and there is a definite capacity which represents the most economical degree of concentration of the wires in the cross-section of the cable. The spacing and inductance of loading coils presents another problem in balanced costs. Even in the case of wooden poles we make use of economic cost studies.

The length of life of a pole depends upon a variety of factors, the most important of which are the character of the timber; whether or not a preservative treatment is employed and, if so, the nature of the treatment; the local climatic and soil conditions and the original size of the pole.

The strength of a pole varies with the cube of the diameter of the sound wood at the weakest section. If the original size of the pole is only slightly more than the critical size at which replacement should be made, the life of the pole will be very short, as decay will reduce the size at the ground line to the critical size within a few years. On the other hand, a pole of huge size at the ground line would have a very long life before rotting sufficiently to require replacement, but the first cost of so stout a pole might readily be so great that its annual cost would exceed that of a smaller and cheaper pole. In our specifications for poles we have constantly to bear in mind that the elimination of poles containing timber defects of one kind or another means that we are adding something to the first cost of our poles and the criterion must always be whether or not the elimination of these defects will sufficiently prolong the life of the poles to warrant the increased first cost.

There have now been placed before you several examples of problems occurring in the telephone industry in the solution of which engineering cost studies may be advantageously employed, and, probably, enough has been said to make clear the importance of this form of economic analysis.

Factors Entering into Annual Costs and Their Evaluation

Let us now consider together the principal factors entering into annual cost, and how, in the course of our work, we evaluate them.

The several factors are these:

1. Cost of money.
2. Taxes.
3. Insurance.
4. Depreciation.
5. Current Maintenance.
6. Administration.
7. Operating Costs.

Cost of Money. The operating companies of the Bell System obtain the new money that they use in extensions to their plants from the sale of their capital stock and securities—bonds and notes. Such a return must be paid the investor, by the Company, as will induce a constant flow of new capital into the business. This steady influx of new capital is required because the System can not decline to expand. It is obligated to meet the increasing needs of the public it serves. Its need for new capital is a direct result of public demand for the service it renders. The rates for service which public utilities may charge are regulated by the commissions, but neither the commissions nor the utilities can fix the worth of money. Public utilities must pay the cost of money just as they must pay the cost of labor, poles and other material. No investor can be forced to invest. If the rate is below what money is worth in the general money market, he will keep out. Utility companies must bring their offerings to a general money market and submit them, in open competition, with the offerings of undertakings of every kind requiring capital. There are two ways of getting new money:

1. From investors willing to lend. These are the bond and note holders.
2. From investors willing to become partners in ownership. These are the stockholders.

Not only do stockholders expect a higher return than bond and note holders, but if the stockholders' earnings are insufficient, the bond investor will take his money to some safer market. Taking into account the ratio which must be prudently maintained between funded debt and stock, a proper figure should be obtained as representing the average annual cost of money. This figure should not be confused with the figure that represents a fair rate of return including a margin for surplus and contingencies.

Taxes. Taxes are levied by various governmental bodies, municipal, county, state and federal, on many different bases. In some specific plant problems, taxes have to be computed to meet the conditions of the case at hand but, in general, it is sufficient to employ a percentage charge for taxes based upon the average experience.

Insurance. In the case of buildings, and equipment contained in buildings, an annual cost item to cover insurance should be included.

Depreciation. Depreciation may be defined as the using up of property in service from all causes. These causes include:

(a) Wear and tear, not covered by current repairs.
(b) Obsolescence.
(c) Inadequacy.
(d) Public Requirements.
(e) Extraordinary Casualties.

All telephone property, except land, is subject to deterioration, and the continued consumption of the investment is a part of the cost of the service which must be provided for by charges against earnings. Only a small portion of the plant actually wears out in service. Instances of this are the rotting of poles and the rusting of iron wire, a relatively small amount of which is used in the plant.

On the other hand, it has been the history of the telephone business that enormous amounts of plant have been taken out of service through no defect in their physical condition but either because they had become obsolete through the development of some more economical or efficient type of equipment, or because they had become inadequate to serve the growing needs of the business.

An example of obsolescence is the replacement of antiquated methods of distribution by more modern types. Examples of inadequacy are the replacement of open wires by cable, and the replacement of small cables by larger ones. Examples of public requirement are the abandonment of pole lines and their replacement by underground construction due to road improvements, and the rebuilding of sections of underground conduit due to changes in the grade of streets or to the construction of transit subways. Examples of extraordinary casualties are fires, sleet storms and tornadoes.

The annual charge for depreciation is an amount which, if entered in operating expenses each year during the service life of a unit of plant, would, at the end of that service life, yield a sum equal to the total depreciation of that unit; that is, its first cost in place less the net salvage obtained at its removal. The consumption of capital is a necessary part of the cost of furnishing service and must be provided for by charges against earnings during the life of the property. In arriving at this depreciation charge the best thing we can do is to take our experience of years and look over the whole situation and apply our judgment to it. The value of this judgment depends on the experience, knowledge, ability and integrity of the people who exercise it.

The amount of this charge should be determined for each broad class of plant and it depends upon the average service life and the net salvage value. Net salvage value is gross salvage value minus cost of removal, and takes into consideration both value for reuse and junk value. For instance, the net salvage value of station apparatus is relatively high because a large part of the equipment can be reused in another location. In other cases, such as iron wire, the net salvage value may be a minus quantity, as there is little or nothing to offset the cost of removal.

Current Maintenance. Current maintenance charges comprise the cost of repairs, rearrangements and changes necessary to keep the plant in an efficient operating condition during its service life. In cost studies, current maintenance charges should be derived from experience and expressed, generally, on a unit of plant basis, as. for example, per pole, per mile of wire, per foot of cable, or per station, according to the kind of plant being considered. Generally speaking, they bear no direct relation to first cost of plant as other annual charges do.

For this reason, when comparing the annual costs of two or more plant units of different sizes or types, an incorrect result would be obtained if maintenance charges were expressed as a percentage of the first cost.

However, for comparative cost studies of average plant, maintained under average conditions, it is sometimes within the precision of the study to employ figures expressed as a percentage of the first cost, provided the figures were derived from the cost of maintaining average plant where average conditions were known to obtain.

Administration. In certain cost studies, a small allowance is usually made to cover that portion of the salaries and expenses of the general officials of the Company which is fairly chargeable to the administration of the plant.

Operating Costs. In certain classes of engineering cost studies, comparisons may involve the situation where one type of plant costs initially more than an alternative type, but permits savings to be made in the daily operating labor which may or may not offset the additional first cost. In such cases, to obtain a true comparison, the operating labor costs under each plan must be combined with the total annual charges which are applied to the first costs of the respective plant quantities.

Present Worths

Engineering cost studies frequently involve a balance between plant installed at the present time and plant installed at some future time. An example of this would be the comparison of a pole whose life was to be extended by attaching it to a stub after (say) 15 years, with a stouter and more expensive pole installed at present or with a pole to which preservative treatment was applied prior to its installation.

In such cases it is not sufficient to compare annual costs which are to be incurred at different times without reducing them to a basis upon which they can properly be compared. If a given amount is required to be expended at some future time, it obviously requires a smaller sum at present in hand to meet this obligation if the fixed time is far distant than if it is in the immediate future.

Let us picture ourselves at the end of the year 1924. If an annual charge of $1,000 is to be paid each year for the 5 years beginning January 1, 1925 and ending December 31, 1929, there will be required, to provide for these five $1,000 payments, the sum of $4,100, in hand, assuming that interest is compounded annually at 7 per cent. On the other hand, if these five annual payments of $1,000 each instead of beginning in 1925 were to begin ten years later, that is, if they were to run from January 1, 1935 to the end of 1939, we should require, in hand, $2,081, that is, only about half as much.

To compare, upon a fair basis, expenditures that have to be made at different times, it is customary, as has been done in the preceding example, to reduce these different expenditures to their "Present Worths," or the equivalent in equated or accumulated annual charges.

Summary

From all that has been said, it becomes evident that, whenever a specific addition is made to a growing plant, we are, to a greater or less extent, committing ourselves to a definite programme for relieving, reinforcing or replacing it at some future time in order most economically to provide for the requirements of growth.

The underlying thought, which can not be overemphasized, is so to plan the plant that, as far as practicable, it will serve for its full life, and require no wholesale changes involving the abandonment of substantial portions of the installation. While the design should be based upon the best estimates of future growth that are obtainable, it must be recognized that the most carefully designed plant layouts employing the best possible estimates of growth, may not always meet the ultimate requirements of flexibility. The chances of a comprehensive plan not fitting in with future development can, however, be reduced to a minimum by thoughtful initial planning.

Generally speaking, our distributing plant layout, once it is established, can not readily nor economically be materially changed. Consequently, if it is not sufficiently flexible in the fundamentals of its design to meet reasonable future possibilities, it may affect adversely the carrying out of proper and economical relief measures, or may require abnormally early reconstruction or replacement. It is very desirable, therefore, always to keep in mind, in any plant layout work, the progressive relief steps which are likely to be required to meet the changing conditions affecting the service requirements. Whenever plant is moved, or taken out of service, property loss is realized. Certain expenditures for these purposes represent the most economical way of conducting the business. But it is of the utmost importance that they should always be incurred along the line of maximum economy, which means that behind every plant addition must be engineering cost studies to assist in furnishing the answers to the three questions:

Why do it at all?
Why do it now?
Why do it this way?

But it must always be borne in mind that these studies do not and can not, in themselves, constitute the sole criterion for determining what should be done. They are, at the best, only an aid, guide and check to be utilized, within their limitations, in arriving at conclusions that must, in the last analysis, rest upon seasoned judgment and experience.

Nevertheless, so great do we find the importance of these engineering cost studies in our work, and so great must be their importance in the engineering of any other kind of growing plant, that the question might be raised whether, in courses of engineering instruction, a few hours at least could not advantageously be devoted to acquainting the student with the nature and importance of these economic problems.

The Limitation of the Gain of Two-Way Telephone Repeaters by Impedance Irregularities

By GEORGE CRISSON

Introduction

Because of the fact that it is a difficult and expensive matter to build and maintain the high grade circuits that are required for modern long distance telephone transmission with repeaters, many workers in this field have attempted to devise some form of two-way repeater which would be able to give as large a gain as desired without singing or poor quality due to irregularities existing in the lines. They have thought that if such a repeater could be constructed it would permit the use of lines less carefully built and, therefore, cheaper than are at present required, and that fewer repeaters would be required because larger gains could be obtained at each repeater.

As a matter of fact the irregularities in the lines have a very important effect and control, to a great extent, the repeater gains which can be used whenever a telephone circuit is arranged so as to be capable of transmitting in both directions over a single pair of wires with constant efficiency.

It is the object of this paper to explain, in a very simple way, why this is true. To do this the phenomenon of electrical reflection is first made clear. Then a two-way repeater system is introduced and the effects of reflection upon this system are explained. After mentioning several of the types of repeaters which have been used successfully, the paper concludes with an explanation of the fallacies underlying a number of schemes which have been proposed from time to time by various inventors.

Reflection in Telephone Lines

Whenever discontinuities or irregularities exist in telephone circuits, reflection of a certain part of the speech wave takes place at each irregularity. In order to appreciate why it is that irregularities in two-wire telephone circuits affect very greatly the amount of repeater gain which can be secured whenever two-way operation is desired, it is first necessary to obtain a clear picture of why it is that reflections take place at irregularities.

Fig. 1 represents an infinite ideal telephone line without repeaters. If such a line is non-loaded or continuously loaded each part of it is exactly like every other part having the same length. If the line is loaded with coils then each loading section is exactly like every other loading section.

When a telephone transmitter or other signaling device *A* acts upon such a line it causes a wave to travel over the line away from

Fig. 1

the source. If the line includes resistance or other losses this wave gradually becomes smaller until it is too weak to be detected but no portion of the wave returns to the source after once leaving it.

If some portion of the line differs in its electrical makeup from other portions of the line it constitutes an irregularity and interferes with the passage of the wave.

Fig. 2 shows a line exactly like that of Fig. 1 except that an irregularity *B* has been introduced. This irregularity has been shown

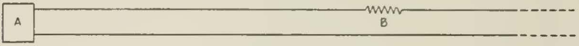


Fig. 2

as a series resistance though any other departure from the regular electrical structure of the line would produce similar effects.

When a wave encounters such an irregularity, it splits into two parts one of which continues in the original direction of propagation along the line while the other is propagated in the opposite direction toward the source.

In order to understand this phenomenon, which is called reflection, imagine that a wave is traversing the line from left to right. As it passes the point *B* a current flows through the series impedance which constitutes the irregularity and this causes a drop of potential through the impedance. Obviously, this changes the state of affairs as there is now a sudden alteration in the voltage across the line as the wave passes the irregularity whereas there is no such alteration without the irregularity.

Suppose that for the impedance element we substitute the output terminals of a generator which has a negligible impedance and arrange the generator so that it is excited by the wave traveling over the line but that the excitation is not affected by the voltage set up by the generator itself. Such an arrangement is shown in Fig. 3. The arrangement for exciting the generator is supposed not to require an appreciable amount of power or to constitute an irregularity. This generator then resembles the series impedance of Fig. 2 in that it produces no disturbance in the line when no waves are passing but as soon as a wave arrives the generator becomes active and produces a

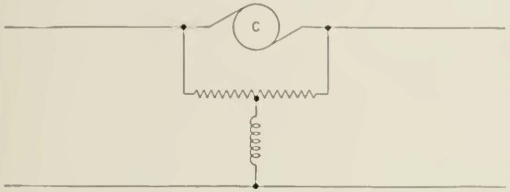


Fig. 3

voltage in series with the line. By proper adjustment of the exciting mechanism of the generator the voltage across its output terminals can be made just equal to the disturbance produced by the impedance element at *B* in Fig. 2 and so exactly reproduce the effects of the irregularity. In order to do this the generator might have to absorb energy from the wave passing over the line instead of giving it out, but it would establish the desired voltage relations.

Now as the generator has no appreciable impedance the wave passes through it without interference but the e.m.f. which it sets up obviously sends out waves in each direction from the generator.

On the right of the irregularity will be found one wave made up of the original undisturbed wave combined with that from the generator and traveling onward in the original direction. The combined wave will usually be smaller than the original wave though it might under some circumstances be larger and its shape might or might not be altered depending upon the nature of the irregularity and the character of the line.

On the left of the irregularity will be found the original wave traveling from left to right and the reflected wave traveling from right to left.

By a similar process of reasoning the reflection caused by bridging an impedance across the line at the point *B* can be illustrated. In this case the output terminals of the generator should be bridged across the line and made of very high impedance.

Any departure from the regular structure of the line such as occurs at the junction of two lines of different types or where loading coils have the wrong inductance or are wrongly spaced causes reflections in the manner described above.

Ideal Repeater on an Ideal Line

Fig. 4 shows an ideal telephone circuit consisting of two sections of line L_1 and L_2 which are free from irregularities and are joined by a repeater R. The remote ends of the line sections are connected to terminal apparatus A_1 and A_2 which have impedances which

Fig. 4

smoothly terminate the lines, that is, if either line had originally extended to an infinite distance from the repeater and had been cut to connect it to the terminal apparatus, this apparatus would have the same impedance as the part of the infinite line which was cut off. The construction of the repeater R is limited only by the requirement that if an electric wave arrives at the repeater terminals T_1 or T_2 over either line a similar but larger wave is transmitted from the repeater over the other line. The gain of the repeater determines the relative sizes of the waves arriving at and departing from the repeater.

If now a wave is started at one end of the circuit, for example A_1, it traverses the line L_1 and is absorbed or dissipated in the portion of the repeater connected to the terminal T_1. This wave acts upon the internal mechanism of the repeater in such a way as to send out a larger wave which traverses the line L_2 and is completely dissipated in the terminal apparatus A_2.

Ideal Repeater on a Line Containing Irregularities

Fig. 5 illustrates a line exactly like that of Fig. 4, except that an irregularity B_1 (or B_2) has been introduced into each section. If a

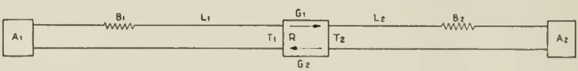


Fig. 5

wave leaves one terminal such as A_1, it traverses the line L_1 eventually arriving at the terminal T_1 of the repeater R with a certain strength. This wave is amplified and transmitted into the line L_2 which it follows until it encounters the irregularity B_2. At B_2 it is partially reflected, one portion returning to the repeater and the other traveling to the terminal A_2 where it is absorbed. The reflected wave passes through the repeater, is amplified and transverses the line L_1 until it encounters the irregularity B_1 where it is again reflected, one part being propagated to the terminal A_1 where it is dissipated, while the other part returns to the repeater and repeats the cycle of amplification and reflection. This action continues indefinitely the wave being reflected alternately from the irregularities B_1 and B_2.

If the total gain in the round trip path is greater than the total loss the wave will be stronger on each arrival at any point in the circuit than on the preceding trip and will continually increase in power until the power limits of the repeater or some other cause prevents a further increase and a steady sing is established. If the gain is less than the loss, the wave will become weaker with each trip from B_1 and B_2 and back until it falls below the strength which can be detected.

Evidently, if the repeater gain is made so great that a steady sing is established, satisfactory telephoning over the circuit will be impossible. Serious quality impairment may occur, however, when the gain is not so great as this. Consequently, when irregularities are present in a line containing repeaters, the repeater gains are necessarily limited.

In the above illustration, it was assumed that two irregularities were present. Serious effects, however, due to the production of echo effects which may be heard by the talker, may be produced by reflection from a single irregularity. Consequently, a single irregularity in the circuit will set a limitation on the repeater gain even though it could not cause singing if a 22-type repeater were used.

From the foregoing explanation, it is evident that the effect of the reflections at the irregularities, which limits the repeater gains, is not dependent upon any special properties of the telephone repeaters. These limitations will necessarily exist with any types of repeater whatsoever which have the property of producing amplification in both directions at the same time.

Effect of Using the Wrong Line Impedance

The discussion will now be extended to show that not only must the lines with which a repeater is to work be smooth, if limitation of the gains is to be avoided, but also the repeaters must be designed to fit lines of one particular type. It has just been shown that reflection takes place if a series or a bridged impedance is inserted in a line. This reflection will take place whether the impedance is inserted at some intermediate point in a line or adjacent to a repeater. Inserting such an impedance adjacent to a repeater would, on account of this reflection, seriously limit the gain which could be produced by the repeater. Now inserting an irregularity adjacent to a repeater amounts to the same thing as substituting a line having a different impedance for the line with which the repeater is designed to function. Since any change in the impedance of a line connected to a repeater away from the impedance with which the repeater is designed to work is equivalent to inserting an irregularity adjacent to the repeater, it is evident that *it is impossible to construct a repeater system whose amplification will be constant in both directions and whose gain will not be limited by irregularities in the lines and by any departure of the line impedance from that for which the repeater is designed.*

Successful Types of Repeaters

Two forms of repeater circuit, the well known 21 and 22 type circuits, have been developed to the point where they have become highly important and successful parts of the telephone plant. These have been so completely described in a paper entitled, "Telephone Repeaters" by Messrs. Gherardi and Jewett,[1] that no further description will be attempted here. It is sufficient to point out that in the case of the 22 type repeater the necessary impedance requirements are met by providing networks which imitate closely the characteristic impedances of the two associated lines. Any departure of the line impedance from the value for which the network was designed or any irregularities in the line or terminal equipment impose limits on the obtainable gain in the manner described above. In the case of the 21 type circuit the impedance requirements are met by putting the repeater between two similar lines whose impedances balance each other.

Another type of repeater circuit, called the booster circuit, was mentioned in the paper just referred to. This circuit does not depend upon impedance balance in the same way as the 21 and 22 type circuits and it is capable of giving two-way amplification but its performance is even more seriously affected by impedance deviations in the lines than the latter circuits. The booster form of repeater circuit has not yet proved useful in a commercial way.

[1] Proceedings of the American Institute of Electrical Engineers, 1919, page 1255.

Devices Employing Voice Controlled Relays

Many different devices aiming to secure the practical equivalent of two-way repeater operation by means of relays (mechanical or thermionic) controlled by the voice currents themselves have been suggested. In these devices the action of the relays is such that when transmission is passing in one direction through a repeater, the transmission in the opposite direction is either wholly or partially blocked. Evidently the gain of such a repeater as this is not limited by impedance irregularities in the lines, since it is really a one-way device during the passage of speech currents.

Repeaters controlled by voice operated devices will not be discussed here further in view of the fact that the principal object of this paper is to treat repeater systems which are truly two-way in their operation.

Other Types of Repeater That Have Been Proposed

Several of the arrangements that have been proposed by inventors who sought unsuccessfully to produce two-way repeaters not subject to limitation by line irregularities will now be described.

1. *Repeaters Involving Balance.* A great many circuits have been devised which involve the principle of balance. These always involve the same fundamental principle as the hybrid coil used in the repeaters now in commercial service though often the arrangement appears quite different. This principle is that the output energy of the amplifier working in one direction, for example, the east bound amplifier, is divided into two parts, one of which is sent into the line east and the other into the corresponding network. The input terminals of the west bound amplifier are so connected that the effect on them of the current entering the line east is opposed by the effect of the current entering the network and consequently the impedances of the line and network must accurately balance each other to keep the output energy of one amplifier out of the input circuit of the other. Sometimes the balance is effected by connecting the line and network into a common electrical circuit and connecting the input terminals of the amplifier to two points of equal potential in this circuit. In other arrangements two fluxes which depend upon the currents entering the line and network are balanced against each other in the core of a special transformer so that a winding connected to the input of the amplifier is not affected.

Usually the impedance of the network equals that of the line, but arrangements are possible and even have certain advantages in which the energy is not equally divided between the line and network and the impedance of the network is either greater than or less than that of the line in a certain ratio.

Through unfamiliarity with the principles involved the inventors sometimes assume that an approximate balance such as might be obtained by using a simple resistance is sufficient to meet all requirements. None of these arrangements, however, can avoid the effects of departures of the line impedance from the values for which the networks are designed nor can they better the performance of the present repeaters in respect to the effects of impedance departures. Usually such circuits are inferior in some important respect to the arrangements now in use.

2. *Circuits using Rectifiers.* In one type of circuit the inventors propose to use rectifiers to prevent the output energy of one amplifier

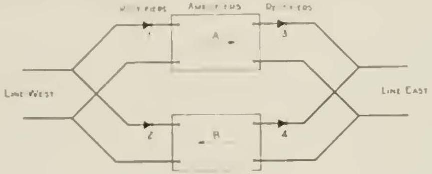


Fig. 6

acting upon the input circuit of the other. A simple diagram illustrating the operation of this scheme is given in Fig. 6. Rectifiers are placed in series with the input and output circuits of both amplifiers and poled in the directions indicated by the arrow heads which point in the direction the rectifier is supposed to permit current to pass. It is argued that the rectifier in the output circuit of each amplifier permits only currents of one polarity to enter the line and that the rectifier in the input circuit of the opposite amplifier is so poled that these output currents cannot pass it into the input circuit and, therefore, singing cannot occur.

If a wave arrives, for example over the line west, the positive half waves pass through the rectifiers 1 and 2 into the input of the east bound and the output of the west bound amplifier respectively. The negative half waves are suppressed by the rectifiers. This is illustrated by Fig. 7 which shows the wave arriving over the line and Fig. 8 which shows the part of the wave which enters the amplifiers.

That portion which reaches the output of the west bound amplifier is lost while the portion which reaches the input of the east bound amplifier, is amplified, and passed on through the rectifier 3 to the line east. If the amplifier were completely distortionless and, therefore, capable of amplifying direct currents and the rectifiers perfect, that is, offering zero resistance to currents in one direction and infinite resist-

Fig. 7

Fig. 8

ance to currents in the opposite direction, the currents transmitted to the line east would have the wave shapes shown in Fig. 8.

As it would be impracticable to make the amplifier amplify the direct-current component of the wave shown in Fig. 8 the amplifier would tend to send out a wave somewhat like that shown in Fig. 9,

Fig. 9

which is the wave of Fig. 8 with the direct component removed. The rectifier 3 then suppresses the negative half waves, finally permitting the wave shown in Fig. 10 to pass to the line east. On account

Fig. 10

of the great distortion involved the quality of speech would be greatly impaired if, indeed, the speech would not be rendered unintelligible.

Assuming, however, that intelligible speech is possible in spite of this distortion, the rectifiers would not prevent singing. Suppose the repeater shown in Fig. 6 to be cut into the line shown in Fig. 5 at R and that waves are arriving from the line west. There are certain line conditions which are practically certain to exist and which would send back reflected waves that would reverse the potential across the line east at the terminals of the repeater, causing impulses to reach the input of the west bound amplifier. These impulses will be amplified and returned to the line west where, if similar conditions exist, they will once more enter the east bound amplifier. If the gains are great enough to offset the losses caused by the rectifiers, the system will sing.

It is, therefore, evident that rectifiers offer no chance for improving on the action of the present types of repeaters because they cause

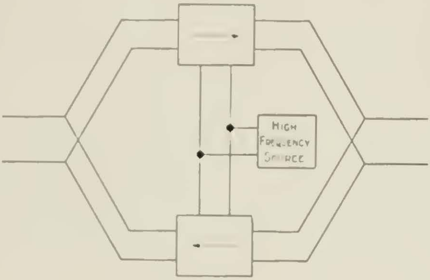


Fig. 11

serious distortion and do not prevent singing except under certain special conditions not likely to be found under practical conditions.

3. *Circuits using High-Frequency Switching.* Another device which is frequently proposed in one form or another is illustrated in Fig. 11. In this case an amplifier is provided for each direction of transmission. These amplifiers are so designed that their amplifying power can be destroyed and restored periodically at high frequency by currents from a suitable source, the amplifier in one direction being active when the other is inactive. The frequency of the controlling currents is above the audible range. In a variation of this scheme a single amplifier is used which is pointed first in one direction and then in the other at a frequency above the audible range. It is argued that since there is amplification in only one direction at any given instant the system cannot sing.

Imagine such a repeater to be inserted in the line at R in Fig. 5, and that voice waves are arriving over the line from A_1. Owing to the nature of the repeater these waves will be cut up into a series of pulses having a frequency equal to that of the controlling current and varying in magnitude according to the shape of the voice wave being transmitted. These pulses will be partially reflected at the irregularity B_2 and part of their energy will return to the repeater. Due to the fact that a finite time is required for the pulses to pass from R to B_2 and back, they are likely to arrive at the right moment to find the amplifier set for amplification in the opposite direction, in which case they will pass through towards A. For a single irregularity, it would be possible to select a frequency such that the pulse would return when the repeater is set against it, but this would require a different frequency for each irregularity which is obviously impossible.

In case the line cannot transmit the high frequency pulses, their energy would be stored in the inductance or capacity of the first elements of the line L_2 and returned to the amplifier when it is in condition to transmit from L_2 to L_1. To avoid the latter objection it has been proposed to employ low pass filters on the output side of each one-way amplifier to convert the high frequency pulses back into ordinary voice waves before passing them into the line, but this obviously defeats the object sought in using the high frequency control of the amplification because each amplifier now receives ordinary voice waves and gives out enlarged copies of them which are subject to the same reflections as if plain one-way amplifiers without the high frequency control had been used.

From these considerations it will readily be seen that repeater systems depending upon high frequency variation of the gain to avoid singing and the necessity for impedance balances are inherently unworkable.

Practises in Telephone Transmission Maintenance Work[1]

By W. H. HARDEN

Synopsis: This paper describes the practical applications of transmission maintenance methods in a telephone system. The methods applicable to toll circuits of various types are first discussed, information being included in this connection on the maintenance of the amplifier circuits involved in telephone repeaters and carrier. Testing methods applicable to the local or exchange area plant are next described, the description including both manual and machine switching systems. The results accomplished in toll and local transmission maintenance work are considered from the standpoint of the kinds of trouble which can be eliminated and the effect which these troubles have on service.

The methods described in the main body of the paper relate particularly to tests of volume efficiency. Certain other transmission maintenance testing methods directly associated with volume efficiency tests are briefly described in Appendix A of the paper.

IT is the purpose of this paper to present a general picture of the practical applications of methods of measuring transmission efficiency in the Bell System which have been developed by study and experience under plant operating conditions. The rapid growth of the telephone industry has made it necessary that these methods be such as to allow them to be applied on a large scale in a systematic and economical manner thereby providing for a quick periodic check of the efficiency of the various types of circuits as they are used in service.

Transmission maintenance can be broadly defined as that maintenance work which is directed primarily towards insuring that the talking efficiencies of the telephone circuits are those for which the circuits are designed. There are, of course, many elements which affect the talking efficiency and various d-c. and a-c. tests are available for checking the electrical characteristics of circuits and equipment to insure that these characteristics are being maintained in accordance with the proper standards. In the final analysis, however, an overall test of the transmission efficiency of the circuit in the condition it is used in service will show at once whether it is giving the loss, or in the case of amplifier circuits, the gain which it should give. Transmission tests, therefore, offer a means whereby many of the electrical characteristics of circuits can be quickly and accurately checked.

In referring to transmission testing apparatus in this paper, four standard types described in previous papers are involved. The first three types listed below were described by Best and the fourth by

[1] Paper presented at the Pacific Coast Convention, *A. I. E. E.*, October, 1924; abstracted in the *Journal*, *A. I. E. E.*, Vol. 43, p. 1121, 1924

Clark.[2] Reference in these papers was also made to the standard oscillators used in supplying the measuring currents for the sets.

1-A *Transmission Measuring Set.* This is an "ear balance" portable set suitable for loop transmission testing only and designed primarily for testing equipment and circuits in the smaller central offices.

3-A *Transmission Measuring Set.* This is a "meter balance" portable set suitable for both loop and straightaway transmission testing and designed primarily for testing circuits and equipment in the larger central offices.

4-A *Transmission Measuring Set.* This is a "meter balance" set suitable for both loop and straightaway transmission testing and designed for permanent installation at the larger toll offices primarily for testing toll circuits.

2-A *Gain Set.* This is a "meter balance" set designed for measuring amplifier gains.

Certain other testing methods in addition to volume efficiency tests are also extensively used in transmission maintenance work and some of the more important of these are briefly discussed in Appendix A of this paper.

Since the routine procedures in testing toll circuits using the above apparatus differ considerably from those followed in the local or exchange area plant, the toll and local practices have been considered separately in the following discussions:

Transmission Tests on Toll Circuits

The importance of having available means for quickly checking the transmission efficiency of toll circuits and of economically maintaining the proper standard of transmission is evident when it is considered that in a plant such as that operated by the Bell System there are at the present time more than 20,000 toll circuits in service. The circuits making up this system are of various types and construction, depending on the service requirements and length, and also upon certain other factors determined by engineering and economical design considerations.

From the standpoint of maintaining transmission efficiency between toll offices, the various types of toll circuits can be divided into three general classes: one, non-repeatered circuits, two, circuits equipped

[2] F. H. Best, "Measuring Methods for Maintaining the Transmission Efficiency of Telephone Circuits," *Journal of the A. I. E. E.*, February, 1924. A. B. Clark, "Telephone Transmission over Long Cable Circuits," *Journal of the A. I. E. E.*, January, 1923.

with telephone repeaters and three, circuits equipped for carrier operation. The latter two classes are alike in many respects as far as the maintenance methods are concerned and both require somewhat more attention than the circuits not equipped with amplifying apparatus. The length and number of repeaters involved are also important factors which must be taken account of in tandem repeater and carrier circuit maintenance. Very long tandem repeater circuits

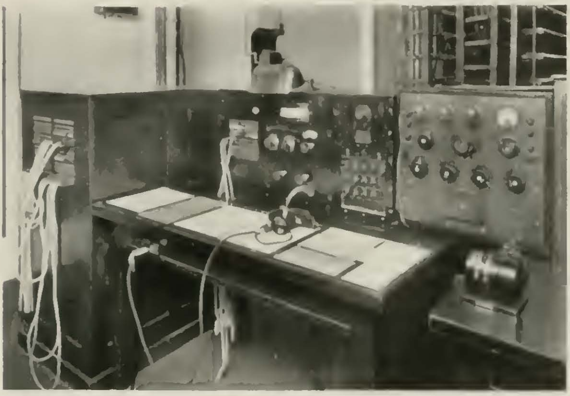


Fig. 1—Illustration of 4-A Transmission Measuring Set and 4-B Oscillator Installed in a Toll Test Room

such, for example, as the long toll cable circuits described by Clark[2] require special maintenance procedures similar in many respects to those required in carrier maintenance.

The 1-A type of transmission measuring set generally used for testing toll circuits may be considered as a toll transmission test desk. Fig. 1 shows a picture of one of the latest models together with an oscillator for supplying the measuring current, installed at a toll office for use in routine testing. The set is provided with trunks to both the toll testboard and toll switchboard, and also with call circuits to toll operators' positions for use in ordering up circuits for test. The electrical measuring circuit is designed so that tests may be made on two toll circuits looped at the distant end, or straightaway on one toll circuit the distant terminal of which terminates in an office also equipped with a transmission measuring set of the same type.

To illustrate the application of this toll transmission test desk, Fig. 2 shows schematically an arrangement of four toll offices having circuits between them of the three general classes non-repeatered, repeatered and carrier. Offices A and D are equipped with transmission measuring sets of the type shown in Fig. 1. A logical testing

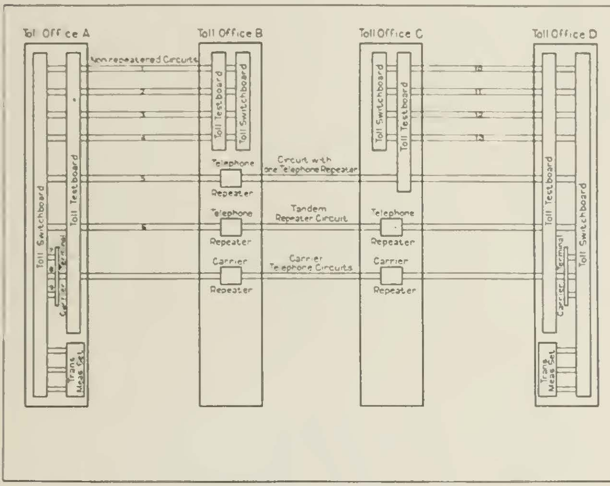


Fig. 2—Schematic Diagram of Typical Toll Circuit Layout to Illustrate General Method of Testing Non-Repeatered, Repeatered and Carrier Circuits

procedure for the arrangement in Fig. 2 is for offices A and D to test the non-repeatered circuits 1 to 4 and 10 to 13 by having them looped two at a time at the distant terminal offices B and C. By "triangulation measurements" on any three circuits in each group, the equivalent of each individual circuit can be readily computed.

For the circuits 5 to 9 extending between offices A and D equipped with telephone repeaters or carrier, straightaway measurements can be made in each direction with the two transmission measuring sets provided. Loop tests could, of course, also be made on the circuits from either office A or D, but this would require cutting the telephone repeaters out of one circuit or having available a non-repeatered or non-carrier circuit, since the gains of the repeaters in the two directions introduce variable factors in the overall equivalents which do not permit triangulation computations to be made. The overall tests on the carrier circuits do not differ in any way from the tests on repeatered or non-repeatered circuits, each carrier channel being tested as a separate circuit through the switchboards. The

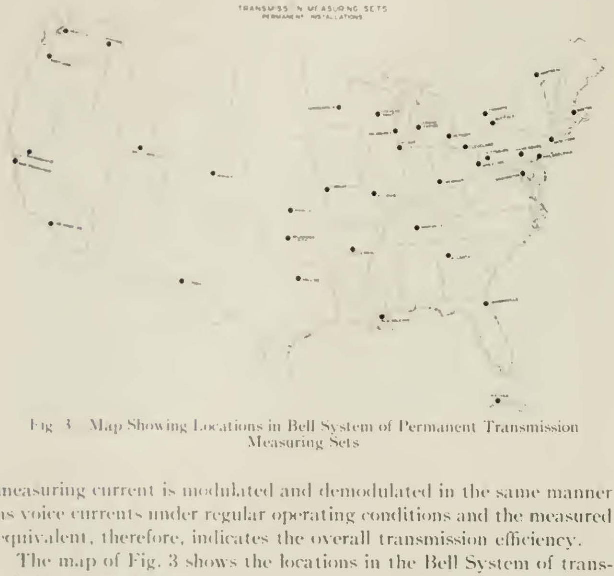


Fig. 3—Map Showing Locations in Bell System of Permanent Transmission Measuring Sets

measuring current is modulated and demodulated in the same manner as voice currents under regular operating conditions and the measured equivalent, therefore, indicates the overall transmission efficiency.

The map of Fig. 3 shows the locations in the Bell System of transmission measuring sets of the general type described above. At a number of the larger toll centers, such as New York and Chicago, where the number of toll circuits to be tested require it, several transmission measuring sets are installed. There are now in operation between 40 and 50 of these sets, making it possible to test all of the longer and more important toll routes in the system. The shorter toll circuits radiating out from the large toll centers are also tested with these same sets. At the smaller offices where fixed transmission measuring sets are not warranted, the toll circuits which cannot be picked up by the larger offices are tested by portable transmission measuring sets of the 1-A or 3-A types in connection with other maintenance work.

One very essential requirement in carrying on a systematic testing program is to have records of the detailed makeup of the toll circuits which give both the circuit layouts and the equipment associated with the circuits. Such a record is valuable, not only in giving the maintenance forces a picture of the circuits and equipment which they are

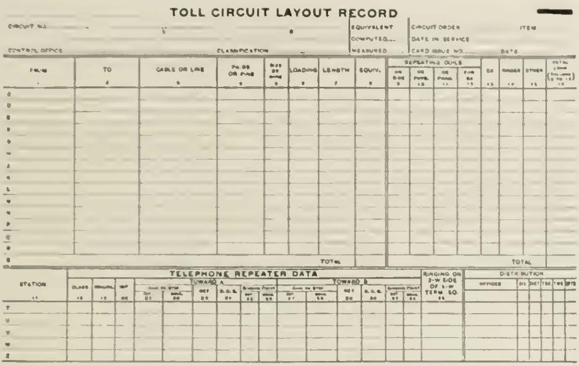


Fig. 4—Sample of a Toll Circuit Layout Record Card

testing, but it also furnishes a means for establishing the transmission standards to which they should work. When transmission tests indicate trouble, this record becomes of particular service in locating and clearing the cause.

Fig. 4 shows a sample of the type of toll circuit layout record card which has proven very satisfactory and is now generally used in the Bell System.

Telephone Repeater and Carrier Maintenance. Voice frequency telephone repeaters were discussed in a paper by Messrs. Gherardi and Jewett[3] and carrier systems in a paper by Messrs. Colpitts and Blackwell.[4] The various arrangements of amplifiers to provide for telephone repeater and for carrier operation as described in these papers make up integral parts of toll circuits and introduce elements

[3] Gherardi and Jewett, "Telephone Repeaters," *Transactions of A. I. E. E.*, 1919, Vol. XXVIII, part 2, pps. 1287 to 1345.

[4] Colpitts and Blackwell, "Carrier Current Telephony and Telegraphy," *Transactions of A. I. E. E.*, 1921, Vol. XL, pps. 205 to 300.

in the circuits which have to be given particular local attention in maintaining the overall transmission efficiency. Since both telephone repeaters and carrier employ the same types of vacuum tubes with very similar arrangements for power supply, the maintenance requirements for the two are much the same. The chief items to be observed in both carrier and repeater maintenance are that the gains specified to give a desired overall transmission equivalent be

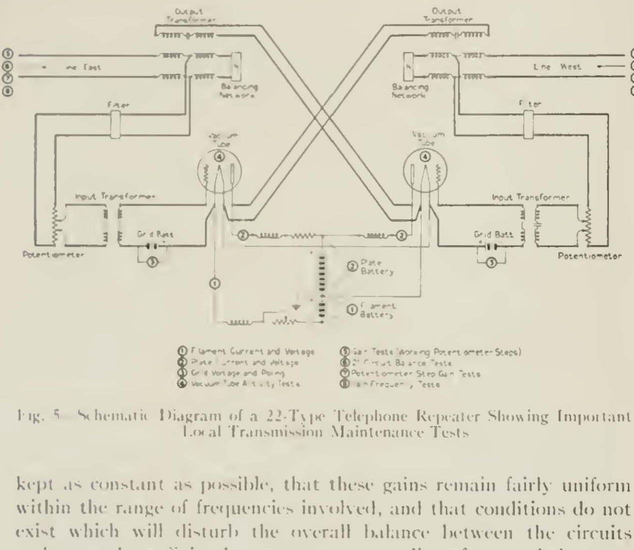
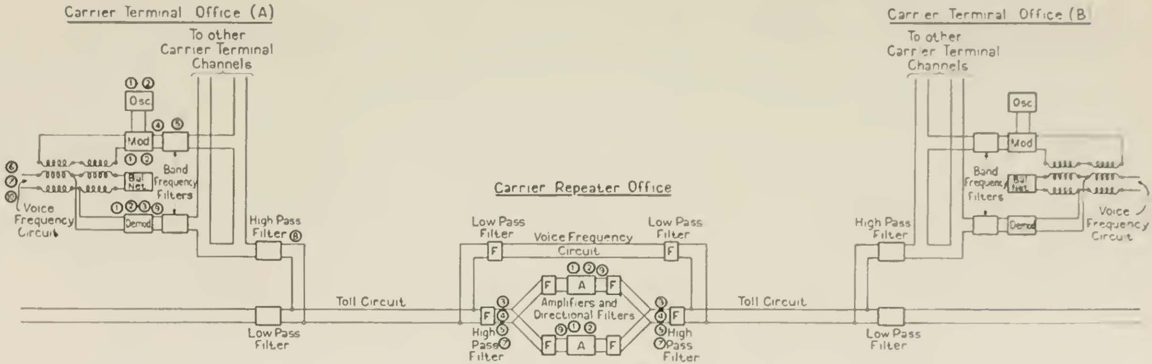


Fig. 5—Schematic Diagram of a 22-Type Telephone Repeater Showing Important Local Transmission Maintenance Tests

kept as constant as possible, that these gains remain fairly uniform within the range of frequencies involved, and that conditions do not exist which will disturb the overall balance between the circuits and networks sufficiently to cause poor quality of transmission.

Considering telephone repeater maintenance, Fig. 5 shows a schematic diagram of a 22 type repeater and indicates the important tests which are made locally to insure that the apparatus is functioning in a satisfactory manner as a part of a toll circuit. The numbers applied to the different tests listed in the figure show approximately the points in the repeater circuit at which the tests are made, the purposes of the tests being evident from their names.

When carrier operation is applied to toll circuits, an additional transmission system is introduced involving the use of currents of higher frequencies than those in the voice range. From a main-



Carrier Terminal Tests

1. Filament, Plate and Grid Battery Tests
2. Vacuum Tube Activity Tests
3. Channel Rectified Received Current Tests
4. Modulator Output
5. Modulator Band Filter Output
6. Channel Loop Gain Tests
7. 21 Circuit Balance Tests on Voice Frequency Circuits

Overall Tests of Complete Carrier System

8. Tests of total Carrier Output Current into Toll Circuit
9. Tests of Carrier Current at Repeater Outputs and finally Rectified Received Current at Distant Terminal
10. Overall Transmission Tests

Carrier Repeater Tests

1. Filament, Plate and Grid Battery Tests
2. Vacuum Tube Activity Tests
3. Gain Tests
4. Potentiometer Step-Gain Tests
5. Gain-Frequency Tests
6. Check of Frequency of Test Oscillator (not shown in figure)
7. High Frequency Singing Tests

9. Output Current on Overall Test of System

Carrier Terminal Tests similar to those listed for Office (A)

Fig. 6—Schematic Diagram of a Carrier Telephone System Showing Important Transmission Maintenance Tests for Carrier Repeaters, Carrier Terminals and Overall

tenance standpoint this means that certain additional testing methods must be employed which will insure the proper generation and transmission of the carrier currents and that the modulation and demodulation of the voice frequency currents is accomplished without distortion or excess loss in overall transmission.

To give a general picture of the more important features involved in the transmission maintenance of carrier systems, Fig. 6 shows a schematic diagram of a carrier layout having one carrier repeater. The particular arrangement shown is for the type *B* system described by Messrs. Colpitts and Blackwell,[4] although the same general maintenance considerations apply to any of the present systems. It will be noted that three series of tests are required, one for the carrier repeaters, one for the carrier terminals and one for the system as a whole. The nature of these various tests and the approximate points in the carrier system where they are applied will be evident from the names and numbers used in the figure.

For both telephone repeaters and carrier systems, provision is made in the regular testing equipment so that the tests can be very quickly applied both as a routine proposition and also when required for trouble location.

Transmission Tests on Exchange Area Circuits

The transmission conditions in the exchange area plant are important not only from the standpoint of insuring good local service but also to insure good toll service, since the local plant forms the terminals of toll connections. The exchange or local plant offers a somewhat different transmission maintenance problem than the toll plant, particularly with respect to the routine testing procedures which must be followed to insure satisfactory transmission. This will be evident when it is considered that in each city and town a complete telephone system is in operation which involves the use of a large number of circuits of various types. There are also in use three general types of telephone switching equipments; manual, panel machine switching, and step-by-step machine switching, and in certain cities combinations of these equipments. It is estimated that at the present time in the Bell System there are in the neighborhood of two and one-half million exchange area circuits, exclusive of subscribers' lines, involving equipment other than contacts and wiring which may directly affect the transmission of speech.

The general classes of exchange area circuits in both manual and machine switching offices, important from a transmission maintenance standpoint, are listed in Table I. The operating features of manual telephone systems are generally well known as are also the features of step-by-step machine switching systems, both having been in use for many years. The panel machine switching system which is a relatively recent development was described in a paper by Messrs. Craft, Morehouse and Charlesworth.[5]

TABLE I

Classification of Circuits in the Exchange Area Plant Important from a Transmission Maintenance Standpoint

MANUAL OFFICES

Local Switchboards	P. B. X. Switchboards	Toll Switchboards	Toll Testboards
Cord circuits	Cord circuits	Cord circuits	Composite set circuits
Operators' circuits	Operators' circuits	Operators' circuits	Composite ringer circuits
Trunk circuits	Trunk circuits	Trunk circuits	Phantom & simplex circuits
Miscl. circuits	Miscl. circuits	Miscl. circuits	Miscl. circuits

Subscribers' loops and sets
Operators' telephone sets

MACHINE SWITCHING OFFICES

Panel	Step by Step
District selectors	Connectors
Incoming selectors	Toll selectors
Trunk circuits	Trunk circuits
Miscl. circuits	Miscl. circuits

Subscribers' loops and sets
Operators' telephone sets for
Special service positions

General classes of exchange area circuits involving equipment other than contacts and wiring which affect telephone transmission.

While it may appear at first hand from the above discussion that transmission testing in the exchange plant is a complicated and expensive matter, this has not proven to be the case. It has been found by experience that the systematic use of transmission measuring sets, following the testing methods which have been developed provides a means for periodically checking transmission conditions with a relatively small amount of testing apparatus and with a small maintenance force. All of the transmission circuits exclusive of subscribers' lines in a 10,000-line central office, either manual or machine switching, can, for example, be completely tested by two men in a

[5] Craft, Morehouse and Charlesworth, "Machine Switching Telephone System for Large Metropolitan Areas," *Journal of the A. I. E. E.*, April, 1923.

period of from two to four weeks, (five and one-half 8-hour days per week assumed) any trouble found being cleared as the testing work is done. The maintenance of the subscribers' lines is not included in this work since it is taken care of by other methods as outlined later.

In order to give a general picture of the application of transmission testing in the exchange telephone plant, a brief discussion of the methods employed in both manual and machine switching systems is given below. In either system the loop method of testing proves

Fig. 7—Illustration of a 3-A Transmission Measuring Set Being Operated in a Manual Office

most satisfactory, that is, one measuring set is used and where both terminals of a circuit are available as in cord circuits, a loop test through the circuit is made. In testing trunk circuits two trunks are looped together at their distant terminals and a measurement made on the two combined.

Transmission Tests on Manual Exchange Area Circuits. In central office, P. B. X. and toll switchboards, the cord circuits and associated operators' circuits are tested by using a portable transmission measuring set, moving this along the boards as required to pick up the cords. Fig. 7 shows a 3-A transmission measuring set being operated at an

A switchboard position. The cords are picked up and plugged directly into the set as shown and measurements made of the loss of both the cord and operator's circuits. Trunk circuit tests are made at the switchboards in the same manner as previously described for loop transmission tests on toll circuits, portable measuring sets such as shown in Fig. 7 generally being employed for this work. Operators' sets are inspected periodically and transmitter and receiver efficiency testing methods are under field trial which provide a means for testing these instruments in central offices. The miscellaneous transmission circuits in an office are tested at the points where they can be most conveniently picked up. The tests on toll test board circuits are made at this board and involve chiefly loop tests on the equipment associated with the toll circuits in the office and tests on the toll line circuits between the toll testboard and toll switchboard.

Transmission Tests on Machine Switching Circuits. The transmission circuits in panel machine switching systems are identical to those in manual systems, while these circuits in step-by-step systems are of a different design but essentially the same as far as transmission losses are concerned. Transmission tests on machine switching circuits are similar to those on manual circuits but involve special methods for picking up the circuits and holding them while the measurements are made. The standard types of transmission measuring sets are used in this work in conjunction with the regular testing equipment provided in the machine switching offices and the methods which have been developed offer a quick and convenient means for making the tests. In manual offices the circuits terminate in jacks or plugs at switchboards where they are readily accessible. In machine switching systems, provision is made for terminating the circuits in jacks at test desks or frames where they can be picked up by patching cords and tested as conveniently as the corresponding types of circuits in manual offices. Machine switching systems offer an important advantage in transmission testing work, particularly in trunk testing, in that the circuits to be tested can be looped automatically by the use of dials or selector test sets, thereby doing away with the necessity for having someone at the distant office complete the loops manually.

In panel machine switching offices the circuits involving transmission equipment corresponding to cord circuits are the "district" and "incoming" selectors. These are tested by setting up the transmission measuring set at the district or incoming frames and connecting the set to the test jacks associated with the circuits. Tests on trunks between manual and panel machine switching offices where both systems are in operation in the same exchange area are generally made from the manual office, the loops being dialed from the *A* switchboard, while trunks between two machine switching offices are tested from the outgoing end of the trunks.

Fig. 8 shows a 3-A transmission measuring set as used in a machine switching office ready for making tests on district selectors. To

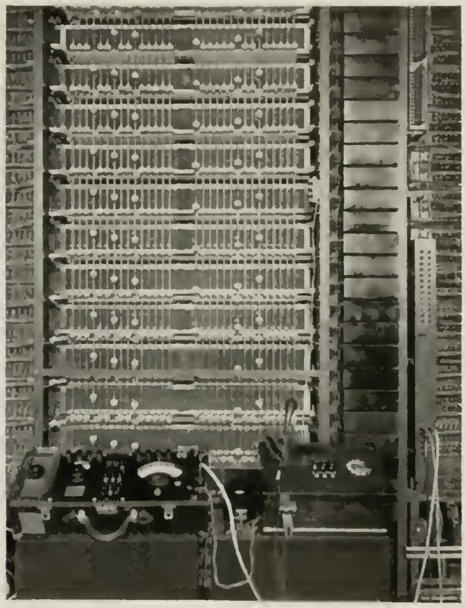
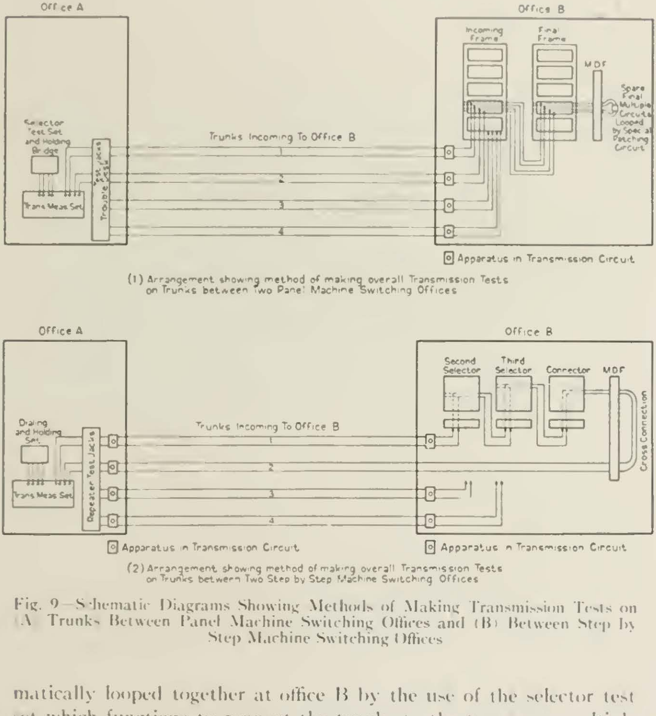


Fig. 8—Illustration of a 3-A Transmission Measuring Set, Set up in a Panel Machine Switching Office for Testing District Selectors

illustrate the general method of testing panel machine switching circuits, the upper diagram of Fig. 9 shows the schematic arrangement for measuring trunks between two panel machine switching offices. The transmission measuring set is located at office A, and connection made to the outgoing end of the trunks to office B through the test jacks at the trouble desk. A standard selector test set used in local maintenance work and a high impedance holding coil are also connected to the trunks through the measuring set, these being used to establish the loop and hold this loop while the tests are made. At office B two spare multiple circuits are cross-connected at the main distributing frame. Any two trunks in the group can then be auto-



(1) Arrangement showing method of making overall Transmission Tests on Trunks between Two Panel Machine Switching Offices

(2) Arrangement showing method of making overall Transmission Tests on Trunks between Two Step by Step Machine Switching Offices

Fig. 9—Schematic Diagrams Showing Methods of Making Transmission Tests on (A) Trunks Between Panel Machine Switching Offices and (B) Between Step by Step Machine Switching Offices

matically looped together at office B by the use of the selector test set which functions to connect the trunks to the two spare multiple circuits previously cross-connected at office B.

In step-by-step machine switching offices the circuits involving transmission equipment corresponding to cord circuits are the connectors. Each connector is provided with a test jack through which connection can be made to a transmission measuring set and the loop completed over a test trunk by dialing. Local selectors do not contain any equipment other than contacts and wiring in the transmission circuits but these can be tested in the same manner as connectors if it is desired to check the wiping contacts and wiring. Toll selectors which involve equipment in the transmission circuit can also be tested in the same manner as connectors. Trunks between manual and machine switching offices can be most conveniently tested from the manual office, the trunk loops being established directly by dialing.

To illustrate the general method of testing step-by-step machine switching circuits, the lower diagram of Fig. 9 shows the schematic circuit arrangement for testing trunks between two machine switching offices. The transmission measuring set is located at office A in a position so that it can be patched to the outgoing trunk repeater test jacks and an arrangement for dialing and holding is connected to the trunks through the measuring set. At office B the apparatus in one trunk is disconnected and this trunk used as a test trunk by cross-connecting it at the main distributing frame to a spare subscriber's multiple terminal. All trunks in the group can then be tested by dialing over them, from office A, the number of this spare terminal at office B which automatically loops them back over the test trunk.

Maintenance of Subscribers' Lines and Stations. The circuits making up subscribers' lines from switchboard to instruments consist simply of pairs of conductors, almost always in cable, with the necessary protective devices. These can be checked by certain d-c. tests described in a recent paper.[6] Equipment is also provided in local test boards for use in making talking transmission tests between the station and the test boards. Accurate machine methods for determining the efficiency of transmitters and receivers have been developed for testing new instruments and instruments returned from service.

General Scheme of Testing Exchange Area Circuits. The plan being followed in the Bell System for systematically checking the transmission conditions of exchange area circuits is to have all offices tested periodically by men equipped with portable transmission measuring sets who travel from office to office. It has been found by experience that after an office has once been tested and any transmission troubles eliminated, it is only necessary thereafter to make transmission tests at infrequent intervals, these subsequent tests serving primarily as a check on the local maintenance conditions.

[6] W. H. Harden, "Electrical Tests and Their Applications in the Maintenance of Telephone Transmission," *Bell System Technical Journal*, July, 1924.

With a testing plan of this kind large areas can be covered by a small traveling force with a small amount of testing equipment. This results in a very economical transmission testing program while at the same time insuring that transmission conditions are maintained satisfactorily.

Fig. 10 shows a typical transmission testing team layout. The team is equipped with an automobile which proves an economical means of transportation between offices and exchange areas and

Fig. 10—Illustration of a Typical Transmission Testing Team Layout

provides a convenient method for carrying the testing equipment. During transportation this equipment is packed in padded trunks which insures against injury. In this particular case the equipment includes, in addition to transmission testing sets and oscillators, other apparatus such as a wheatstone bridge, crosstalk set and noise measuring set so that other maintenance work may be done in connection with transmission testing whenever this is desired.

Results Accomplished

The results accomplished in transmission maintenance work can best be appreciated by considering the kinds of troubles which adversely affect transmission and which can be detected and eliminated by routine testing methods. Consideration is first given to the general causes of troubles which are detrimental to both toll and local transmission, and later the features in this connection more particularly identified with telephone repeaters and carrier systems are discussed.

The different classes of circuits given in Table I are made up of various combinations of the following individual parts:

Repeating Coils	Plugs	
Retardation Coils	Jacks	
Relays	Keys	
Condensers	Heat Coils	
Resistances	Carbons	
Auto-Transformers	Wiring	Switchboard to M. D. F.
Induction Coils		Cross-connection
Loading Coils		Outside
Cords	Transmitters	
	Receivers	

The above parts are combined in various ways to make up the complete operating circuits such as cord circuits, operators' circuits, trunk circuits, etc. Each complete circuit causes a definite normal loss to telephone transmission which must be taken account of in designing the plant to meet the various service requirements. If, however, any of the parts used are defective, if the wrong combinations of parts are used, or if the installation work is not correctly done, excess transmission losses will result which may very seriously affect the transmission when the particular circuits involved are employed in an overall connection.

Classification of Common Types of Troubles. An analysis of a large amount of transmission testing data has made it possible to develop a definite trouble classification which is particularly helpful in transmission maintenance work and which permits the most efficient use of the results in eliminating transmission troubles. Experience has shown that the troubles found can be divided into two general classes, A—troubles which can be detected either by simple d-c. or a-c. tests in connection with the regular day-by-day maintenance work or by transmission measuring sets, and B—troubles which can be detected most readily by transmission measuring sets. The most important troubles in the above classes are as follows:

Class A	Class B
Opens	Electrical Defects
Grounds	Incorrect Wiring
Crosses	Wrong Type of Equipment
Cutouts	Missing Equipment
	High Resistance
	Low Insulation

If, in making transmission tests in a central office, a high percentage of Class A troubles is found the remedy is generally to instigate more rigid local maintenance routines paying particular attention to the type of circuits in which the troubles are located. The percentage of Class B troubles is not as a rule as high as the Class A troubles and experience has shown that when Class B troubles are once eliminated by transmission testing methods only infrequent subsequent tests are required to take care of any additional troubles of this class which may get into the plant.

In determining what constitutes an excess loss, the value of the transmission as well as the practical design and manufacturing considerations to meet operating limits are taken account of. An excess gain is also considered as a trouble on circuits equipped with amplifiers, since this may produce poor quality of transmission which is likely to be more detrimental to service than an excess loss. The value of transmission based on economical design considerations varies, depending on the first cost and annual charge of the particular types of circuits involved. A gain of one TU in the toll plant is generally worth more, for example, than one in the local plant, since it costs more to provide. In transmission maintenance work the cost of making transmission tests and clearing trouble is balanced against the value of the transmission gained for the purpose of establishing economical transmission limits to work to.

Specific Examples of Common Troubles Found and Their Effect on Transmission. Certain kinds of troubles which are detected by transmission measuring sets do not cause excess losses which can be quantitatively measured. Such troubles are, however, readily detected by "ear balance" transmission measuring sets in that they cause noise or scratches and by the "meter balance" sets from fluctuations of the needle of the indicating meter. The most common trouble of this kind is due to cutouts or opens which may be caused by dirty connections, loose connections, improper key and relay adjustments, etc. While not causing a quantitative value of excess loss, this class of trouble is very detrimental to transmission and more serious in many instances than fixed excess losses. Indeterminate troubles of this nature are given an arbitrary excess loss value based on experience.

Considering troubles which give definite losses, the most common kinds are caused by electrical defects in equipment, incorrect wiring of equipment in circuits and wrong types of equipment. The other classes of troubles, such as crosses, high resistances, and low insulation, also generally give measurable excess losses but these are not as common in the plant, since troubles of this nature are more likely to affect the signaling and operation of the circuits and are, therefore, eliminated by the regular maintenance work. Missing equipment will in certain cases cause a gain in transmission but affects the circuits adversely in other ways.

Typical examples of common troubles, with the excess losses which they cause, are given in the following table:

Type of Circuit and Equipment	Cause of Trouble	Approximate Excess Transmission Loss[7]
Repeating coils in cords, incoming trunk circuits, selectors, toll connectors	Electrical defects (Generally short circuited turns)	1.5 to 5.0 TU
	Incorrect wiring (Generally reversed windings)	2.0 to 13.0 TU
Supervisory relays in "A" cord circuits	Electrical defects (Open non inductive winding)	About 2.5 TU
Bridged retardation coils or relays in toll cord circuits, composite sets, connectors and step-by-step repeaters	Electrical defects (Generally short circuited turns)	1.0 to 5.0 TU
Repeating coils on loaded toll switching trunks	Wrong type of equipment incorrect wiring	1.0 to 4.0 TU
Induction coils in operators' telephone sets	Electrical defects. Incorrect wiring	1.0 to 13.0 TU

There are, of course, many other specific types of troubles detected by transmission tests which give definite quantitative losses but the above will serve to illustrate the value of this testing work in eliminating excess losses in a telephone plant.

Maintenance Features Peculiar to Telephone Repeaters and Carrier Systems. The same classification of troubles discussed above applies to repeaters and carrier systems. Amplifier equipment, however, employs certain features which are not common to the more simple telephone circuits and some of the troubles which may occur if the proper maintenance procedures are not followed will seriously affect service. It is for this reason that repeater and carrier installations are provided with special testing equipment which is always available for use either in routine maintenance or in locating and clearing any troubles which may occur in service. Automatic regulating devices are also provided wherever this is practicable in order to reduce to a minimum the amount of manual regulation and maintenance.

[7] W. H. Martin, "The Transmission Unit," *Journal of the A. I. E. E.*, June, 1924; *B. S. T. J.*, Vol. III, p. 400, 1924, C. W. Smith, "Practical Application of Transmission Unit," *B. S. T. J.*, Vol. III, p. 409, 1924.

The important elements in both repeaters and carriers which may directly affect transmission or cause service troubles in other ways are as follows:

Filament Batteries	Potentiometers
Plate Batteries	Filters
Grid Batteries	Transmission Equalizers
Vacuum Tubes	Signaling Equipment
Balancing Equipment	Patching Arrangements

The tests outlined in the main body of the paper aim to insure that the above essential parts of repeater and carrier circuits are functioning properly and that the equipment as a whole is giving the desired results in overall transmission efficiency.

Conclusion

The above discussion of testing methods and the results accomplished indicate how a comprehensive and economical transmission maintenance program can be applied to a telephone plant to check the volume efficiency of the circuits against the established standards. Consideration is continually being given to new testing methods and their applications in order that further improvements in service may be effected and increased economies in testing taken advantage of.

APPENDIX A

Principles of Testing Methods Closely Associated with Transmission Efficiency Tests

Tests of volume efficiency often need to be supplemented by other methods of testing in transmission maintenance work. Transmission efficiency both as regards volume and quality may be seriously affected by noise or crosstalk, and tests for any conditions of this kind are therefore important in maintenance work. Furthermore when efficiency tests show excess losses or unsatisfactory circuit conditions other testing methods prove very valuable in locating the cause.

To illustrate this phase of transmission maintenance the principles of some of the more important testing methods are briefly described below. Two of the tests employ a method very similar to loop transmission testing while others employ the well known "null" method. A special method employing three winding transformers and amplifiers widely used to determine impedance balance conditions between lines and networks is also described. Several methods which involve simply current and voltage measurements have been mentioned in this paper but these are generally well known and therefore require no detailed description.

1. Measurements of Crosstalk

In the circuit shown in Fig. 11, if a-c. power is supplied to a circuit known as the "disturbing" circuit and unbalances exist between this circuit and a second known as the "disturbed" circuit, power will be transferred from one circuit to the other causing crosstalk in the second. A definite power transmission loss therefore takes place between the two circuits which can be measured by a loop transmission test similar to the efficiency tests described in the main body of the paper. An adjustable shunt called a "crosstalk meter" cali-

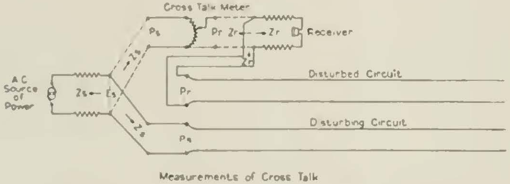
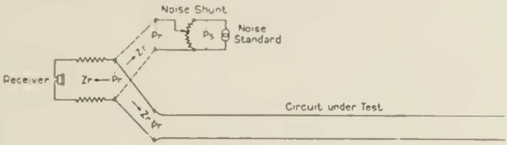


Fig. 11—Diagram Showing Principles of Crosstalk Measurements

brated in either *TU* or in crosstalk units is substituted for the two circuits. With the same power supplied alternately to both the "disturbing" circuit and the meter and with the sending and receiving end impedance conditions as shown, the meter shunt is adjusted until, in the opinion of the observer, the annoyance produced by the tone in the receiver is judged to be equal for the two conditions. The reading of the shunt if there was no distortion of the line crosstalk currents would then give the volume of crosstalk which could be expressed in *TU* as $10 \log_{10} P_s/P_r$ similar to loop transmission testing. However, this relation only holds approximately in practise since the line crosstalk measured is produced by various currents having different phase relations and a certain amount of distortion therefore occurs. The commercial form of crosstalk set now used is equipped to give the approximate impedance relations required and also provides a feature for eliminating the effect of line noise except in the case of one type of measurement which is made on long cable circuits. For practical reasons the results are generally expressed in crosstalk units rather than *TU*.

2. Measurements of Noise

The common method of measuring noise in a telephone circuit is shown in the diagram of Fig. 12. In this test an artificial noise current produced by a generator of constant power P_s called a "noise standard" is substituted for the line noise current. If the two noise currents were exactly alike as regards wave shape and the relative magnitude of the frequencies involved they would produce the same tone in the receiver and their volumes could be made equal by adjustment of the noise shunt. The power ratio, P_r/P_s, as indicated



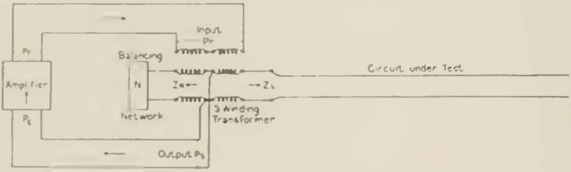
Measurements of Noise

Fig. 12—Diagram Showing Principles of Noise Measurements

by the shunt, would then give a measure of the line noise in terms of the noise standard. This condition, however, is not met with in practise due to differences in wave shape of the two noise currents. For this reason noise measurements are made by adjusting the noise shunt until the interfering effects of the noise on the line and from the shunt are judged to be the same for which condition the power supplied to the receiving network by the noise standard is not necessarily the same as that supplied by the line. The receiving end impedances however, are kept as nearly alike as practicable to prevent reflection losses.

3. Measurements of Line-Network Balance (21-Circuit Balance Test)

The testing arrangement of Fig. 13 shows the principle of the 21-circuit balance test referred to in the main body of the paper in connection with telephone repeater and carrier maintenance. In this test the gain of an amplifier calibrated in *TU* is used to compensate for the loss through a three winding transformer or output coil of a telephone repeater. If the impedances of the balancing network and line were exactly alike at all frequencies, *i.e.*, $Z_n = Z_L$, and no other unbalances existed in the circuit none of the power supplied by the amplifier to the input of the three-winding transformer would be transferred to the output, *i.e.*, the power ratio P_s/P_r would be infinity. However, this ideal condition cannot be produced in practise so that there is always a finite power loss between the input and output of the transformer which can be measured approximately by the gain of an amplifier calibrated in *TU*. An internal



Measurements of Impedance Balance Between Lines and Networks
(21-Circuit Tests on Telephone Repeaters and Carrier)

Fig. 13—Diagram Showing Principles of 21 Circuit Balance Tests

path for currents which may produce "singing" or a sustained tone is established if the gain of the amplifier P_r/P_s is greater than the loss P_s/P_r through the three-winding transformer. As unbalances between network and line become greater the loss through the three-winding transformer becomes less thereby requiring less gain in the amplifier to produce a "singing" condition. It should be noted in this connection that to produce the condition described above exactly, the current received around the "singing" path must be in phase with the starting current. In practise this condition obtains sufficiently accurately so that the gain of the amplifier required to produce "singing" gives an approximate measure of the impedance balance between line and network.

4. Measurements of Resistance, Reactance and Impedance

Diagram (a) of Fig. 14 shows the wheatstone bridge circuit for d-c. resistance measurements. It is unnecessary to describe the well known principles of this bridge but mention is made of it here in view of its importance and use in telephone maintenance work. It supplies an indispensable method of measurement for certain trouble locations, such as crosses and grounds and embodies the fundamental principles of all null tests.

Diagram (b) of Fig. 14 gives a bridge circuit for measuring impedance, the particular arrangement shown being for measurements of impedances having inductive reactance. The bridge measurements express impedance in terms of its resistance component and equivalent inductance or capacity. In measuring an impedance having inductive reactance at any frequency, f, for example, a balance gives $R = R_x$ and $L = L_x$. At the frequency f, the effective resistance is given directly by the value of R and the reactance by the relation, $2\pi f L$.

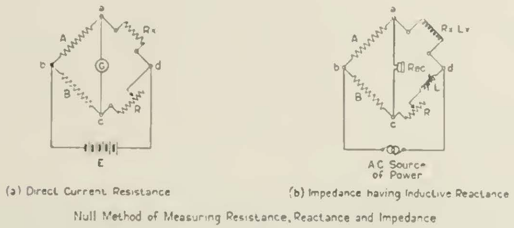


Fig. 14—Diagrams Showing Principles of Null Methods for Measuring Resistance, Reactance and Impedance

The impedance is the vectorial sum of these two or $\sqrt{R^2+(2\pi f L)^2}$. In maintenance work involving impedance measurements as will be noted in the next testing method described, the effective resistance component and the equivalent inductance are generally used directly without combining.

5. Measurements of Line Impedance and Location of Impedance Irregularities

Fig. 15 shows a telephone circuit connected to a bridge and terminated at its distant end in characteristic impedance. If the circuit has approximately uniform impedance throughout its length the resistance and equivalent inductance curves of this impedance within a range of frequencies will be fairly smooth as indicated by *A* and *C* of the figure. The curves are not perfectly smooth since it is not practicable to construct the line for perfect impedance uniformity. If at some point in the circuit an irregularity is present such as an omitted loading coil, an inserted length of line of different construction, etc., which changes the impedance, this will produce a periodic change in the resistance and inductance curves *A* and *C* such as shown by Curve *B*. Curve *C* will be changed in the same way as Curve *A* but for simplification this is not shown on the diagram.

The change in impedance in the circuit reflects some of the current sent out back to the sending end where it adds to or subtracts from the sending current depending on the phase relations of the two currents at any particular frequency. Since impedance equals E/I its value changes as the value of I changes. This is made use of in line impedance measuring work to give a location of impedance irregularities which may exist somewhere in the line.

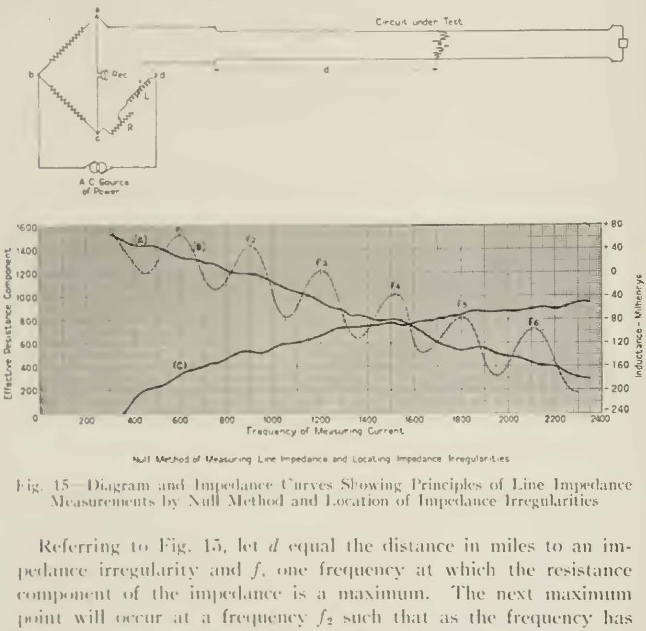


Fig. 15—Diagram and Impedance Curves Showing Principles of Line Impedance Measurements by Null Method and Location of Impedance Irregularities

Referring to Fig. 15, let d equal the distance in miles to an impedance irregularity and f_1 one frequency at which the resistance component of the impedance is a maximum. The next maximum point will occur at a frequency f_2 such that as the frequency has been increased, one complete wave length is added in the distance traveled by the reflected current. Maximum points at f_3, f_4, etc., occur in the same way as the frequency is increased. Considering the two values f_1 and f_2 let

V = velocity of current in miles per second

W_1 = wave length at frequency f_1

W_2 = wave length at frequency f_2

N = number of wave lengths in distance traveled

by reflected current or $2d$.

At frequency f_1 then,

$$N = \frac{2d}{W_1}$$

and at f_2,

$$N + 1 = \frac{2d}{W_2}$$

also at f_1,

$$W_1 = V/f_1$$

and at f_2,

$$W_2 = V/f_2$$

Substituting above

$$N = \frac{2df_1}{V} \text{ and}$$

$$N + 1 = \frac{2df_2}{V}$$

Subtracting,

$$1 = \frac{2df_2}{V} - \frac{2df_1}{V} \text{ or}$$

$$d = \frac{V}{2(f_2 - f_1)}$$

which is the distance in miles from the sending end of the circuit to the point of impedance irregularity. The velocity of propogation V is not exactly constant within the entire frequency range but does not vary sufficiently to materially effect the accuracy of impedance trouble locations by this method.

Mutual Inductance in Wave Filters with an Introduction on Filter Design

By K. S. JOHNSON and T. E. SHEA

PART I

General Principles of Wave Filter Design

Principles of Generalized Dissymmetrical Networks. We shall consider first the impedance and propagation characteristics of certain generalized networks. It can be shown that *any passive network having one pair of input and one pair of output terminals may, at any frequency, be completely and adequately represented by an equivalent T or π net-*

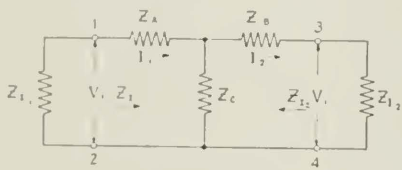


Fig. 1—Generalized Dissymmetrical T Network Connected to Impedances Equal to Its Image Impedances

work.[1] The impedance and propagation characteristics of any such network may be expressed in terms of its equivalent T or π network. These characteristics are defined by (1) the *image impedances*, and (2) the *transfer constant*, the latter including the *attenuation constant*[2] and the *phase constant*.[2] In the case of a symmetrical network, the image impedances and the transfer constant are, respectively, the *iterative impedances* (or *characteristic impedances*) and the *propagation constant* employed by Campbell, Zobel, and others. The terms involved will be subsequently defined.

Consider the dissymmetrical T network of Fig. 1. If the 3–4 terminals of the T network are connected to an impedance Z_{I_2}, the

[1] Campbell, G. A., "Cisoidal Oscillations," *Transactions A. I. E. E.*, (1911), Vol. XXX, Part II, pp. 873–909.

The T and π networks referred to above are sometimes called *star* (Y) and *delta* (Δ) networks, respectively.

[2] The real and imaginary parts of the transfer constant have been called by Zobel, the *attenuation constant* and the *angular constant*, respectively. See Bibliography 13.

impedance looking into the T network at the $1-2$ terminals will be

$$Z_{1-2} = Z_A + \frac{Z_C(Z_B+Z_{I_2})}{Z_C+Z_B+Z_{I_2}}. \tag{1}$$

Similarly, if the $1-2$ terminals of the T network are connected to an impedance Z_{I_1}, the impedance looking into the $3-4$ terminals of the T network will be

$$Z_{3-4} = Z_B + \frac{Z_C(Z_A+Z_{I_1})}{Z_C+Z_A+Z_{I_1}}. \tag{2}$$

If Z_{1-2} is equal to the terminal impedance Z_{I_1}, and if, similarly, Z_{3-4} is equal to the terminal impedance Z_{I_2}, the network will then be terminated in such a way that, at either junction ($1-2$ or $3-4$), the impedance in the two directions is the same. In other words, at each junction point, the impedance looking in one direction is the *image* of the impedance looking in the opposite direction. Under these conditions Z_{I_1} and Z_{I_2} are called the *image impedances* of the T network. If equations (1) and (2) are solved explicitly for Z_{I_1} and Z_{I_2}, the following expressions are obtained:

$$Z_{I_1} = \sqrt{\frac{(Z_A+Z_C)(Z_AZ_B+Z_AZ_C+Z_BZ_C)}{(Z_B+Z_C)}}, \tag{3}$$

$$Z_{I_2} = \sqrt{\frac{(Z_B+Z_C)(Z_AZ_B+Z_AZ_C+Z_BZ_C)}{(Z_A+Z_C)}}. \tag{4}$$

If Z_{oc} is the impedance looking into one end of the network with the distant end open-circuited, and if Z_{sc} is the corresponding impedance with the distant end short-circuited, it may be shown that the image impedance at either end of the network is the geometric mean of Z_{oc} and Z_{sc}. What is here termed the image impedance is, therefore, equivalent to what Kennelly has called the *surge impedance*.[3]

The propagation characteristics of a dissymmetrical network may be completely expressed in terms of the transfer constant. The transfer constant of any structure may be defined as one-half the natural logarithm of the vector ratio of the steady-state vector volt-amperes entering and leaving the network when the latter is terminated in its image impedances. The ratio is determined by dividing the value of the vector volt-amperes at the transmitting end of the network by the value of the vector volt-amperes at the receiving end.

[3] There is at present lack of common agreement as to the basis of definition of this term, and it is often defined upon the basis, not of open and short-circuit impedances, but of a uniform recurrent line (See A. I. E. E. Standardization Rule 12054, edition of 1922). The formulae derived by the two methods are not equivament in the case of dissymmetrical networks.

The real part of the transfer constant, that is, the *attenuation constant*, is expressed by the above definition in *napiers* or *hyperbolic radians* and the imaginary part, that is, the *phase constant*, is expressed in circular *radians*. The practical unit of attenuation here

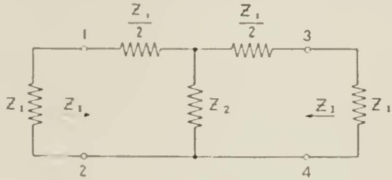


Fig. 2—Generalized Symmetrical *T* Network Connected to Impedances Equal to Its Image Impedances

used is the *transmission unit*[4] (1 $TU = .11513$ napier). It can be demonstrated that the transfer constant, θ, of the *T* network shown in Fig. 1 is

$$\theta = \tanh^{-1}\sqrt{\frac{Z_{sc}}{Z_{oc}}} = \tanh^{-1}\sqrt{\frac{Z_AZ_B+Z_AZ_C+Z_BZ_C}{(Z_A+Z_C)(Z_B+Z_C)}}$$
$$= \cosh^{-1}\sqrt{\frac{(Z_A+Z_C)(Z_B+Z_C)}{Z_C^2}}, \tag{5}$$

in which Z_{oc} and Z_{sc} are, as previously defined, the open and short-circuit impedances of the network. The ratio Z_{sc}/Z_{oc} is the same at both ends of any passive network.

Principles of Generalized Symmetrical Networks. Consider now the impedance and propagation characteristics of the generalized symmetrical structure shown in Fig. 2. On account of the symmetry of the structure, the image impedances at both ends are identical, and from equation (3) or (4) their value may be shown[5] to be

$$Z_I = \sqrt{Z_1Z_2\left(1+\frac{Z_1}{4Z_2}\right)}. \tag{6}$$

In the case of a symmetrical *T* structure, such as is shown in Fig. 2, the impedance Z_I is called the *mid-series image impedance*. The significance of this term will be evident, if the series-shunt type of

[4] W. H. Martin, "The Transmission Unit and Telephone Transmission Reference System," *Bell Sys. Tech. Jour.*, July, 1924, *Jour. A. I. E. E.*, Vol. 43, p. 504, 1924.

[5] Zobel, O. J., "Theory and Design of Uniform and Composite Electric Wave-Filters," *Bell Syst. Tech. Jour.*, Jan., 1923.

structure shown in Fig. 3 is regarded as made up of symmetrical T networks or sections, the junctions of which occur at the mid-points of the series arms.

Suppose now that the structure of Fig. 3 is considered to be made up of symmetrical π networks, or sections, each of which is represented

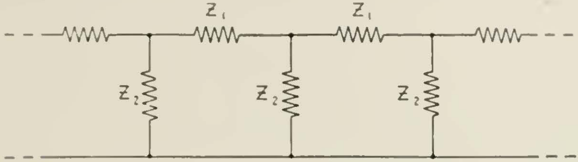


Fig. 3—Generalized Recurrent Series-Shunt Network

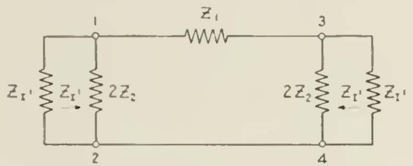


Fig. 4—Generalized Symmetrical π Network Connected to Impedances Equal to Its Image Impedances

as in Fig. 4. By methods similar to those employed for the T network of Fig. 2 it can be shown[5] that the image impedance of the generalized π network of Fig. 4 is given by

$$Z_{I'} = \sqrt{\frac{Z_1 Z_2}{1+\frac{Z_1}{4Z_2}}}. \tag{7}$$

In this symmetrical structure the image impedance is called the *mid-shunt image impedance.*

The image transfer constant of either a T or a π symmetrical structure is[5]

$$\theta = A + jB = 2 \sinh^{-1}\sqrt{\frac{Z_1}{4Z_2}} = \cosh^{-1}\left(1+\frac{Z_1}{2Z_2}\right). \tag{8}$$

In discussing the generalized networks of Figs. 1, 2 and 4, it has been assumed that the networks were terminated in their respective image impedances. In practical cases, filters must be designed to work between impedances which are, in general, not exactly equal to their image impedances at more than one or a few frequencies. For a generalized structure, such as that of Fig. 1, operating between a *sending-end impedance* Z_S and a *receiving-end impedance* Z_R, the current in Z_R, for an electromotive force acting in Z_S, is

$$I_R = \frac{E}{Z_S+Z_R} \times \frac{Z_S+Z_R}{\sqrt{4Z_SZ_R}} \times \frac{\sqrt{4Z_{I_1}Z_S}}{Z_{I_1}+Z_S} \times \frac{\sqrt{4Z_{I_2}Z_R}}{Z_{I_2}+Z_R} \times \epsilon^{-\theta} \times \frac{1}{1-\dfrac{Z_{I_2}-Z_R}{Z_{I_2}+Z_R} \times \dfrac{Z_{I_1}-Z_S}{Z_{I_1}+Z_S} \times \epsilon^{-2\theta}}. \tag{9}$$

Since $E/(Z_S+Z_R)$ is the current $(I_{R'})$ which would flow if the generalized T network were not inserted in the circuit, the ratio of the received current, *with* and *without* the network in the circuit, may be expressed by the relation

$$\frac{I_R}{I_{R'}} = \left(\frac{Z_S+Z_R}{\sqrt{4Z_SZ_R}}\right)\left(\frac{\sqrt{4Z_{I_1}Z_S}}{Z_{I_1}+Z_S}\right)\left(\frac{\sqrt{4Z_{I_2}Z_R}}{Z_{I_2}+Z_R}\right) \times \epsilon^{-\theta} \times \frac{1}{1-\left(\dfrac{Z_{I_2}-Z_R}{Z_{I_2}+Z_R}\right)\left(\dfrac{Z_{I_1}-Z_S}{Z_{I_1}+Z_S}\right)\epsilon^{-2\theta}}. \tag{10}$$

In general, the electromotive force does not act through a simple sending-end impedance Z_S but through some complex circuit. The current ratio $(I_R/I_{R'})$ will, however, be the same in either case. The principle underlying this fact is known as *Thévenin's Theorem*.[6]

The absolute magnitude of the current ratio, $|I_R/I_{R'}|$, is a measure of the *transmission loss* caused by the introduction of the network. The transmission loss may be expressed in terms of transmission units (TU) by aid of the following relation

$$TU = 20 \log_{10} \left|\frac{I_{R'}}{I_R}\right|. \tag{11}$$

Reference to equation (10) shows that the transmission loss caused by the introduction of any network is composed of five factors. The first three factors of this equation are all of the same general type with the exception that the first of the three is reciprocal in nature to the other two. These two latter factors have been called *reflection factors* and determine the *reflection losses* which exist between the impedances involved. The fourth factor is the *transfer factor* and expresses the current ratio which corresponds to the transfer con-

[6] Casper, W. L., "Telephone Transformers," *Transactions A. I. E. E.*, March, 1924, p. 1. Thévenin, M. L., "Sur un Nouveau Théorème d'Electricité Dynamique," *Comptes Rendus*, vol. 97, p. 159, 1883.

stant. The last factor has been called the *interaction factor*. The value of the reflection factor is evidently a function simply of the *ratio* of the impedances involved, while the absolute value of the transfer factor is ϵ^{-A} where A is the real portion of the transfer constant and hence is the attenuation constant. The value of the interaction factor is seen to be unity either when $Z_{I_2} = Z_R$ or when $Z_{I_1} = Z_S$. It also approaches unity if the value of Θ is sufficiently large.

In the case of a symmetrical structure, such as is shown in Fig. 2, or Fig. 4, $Z_{I_1} = Z_{I_2} = Z_I$ and equation (10) reduces to

$$\frac{I_R}{I_{R'}} = \left(\frac{Z_S+Z_R}{\sqrt{4Z_S Z_R}}\right)\left(\frac{\sqrt{4Z_I Z_S}}{Z_I+Z_S}\right)\left(\frac{\sqrt{4Z_I Z_R}}{Z_I+Z_R}\right) \times \epsilon^{-\theta} \times \frac{1}{1-\left(\frac{Z_I-Z_R}{Z_I+Z_R}\right)\left(\frac{Z_I-Z_S}{Z_I+Z_S}\right)\epsilon^{-2\theta}}. \tag{12}$$

If the structure is symmetrical, and if, furthermore, the sending-end impedance Z_S is equal to the receiving-end impedance Z_R, equation (12) becomes

$$\frac{I_R}{I_{R'}} = \epsilon^{-\theta} \times \frac{4Z_I Z_R}{(Z_I+Z_R)^2} \times \frac{1}{1-\left(\frac{Z_I-Z_R}{Z_I+Z_R}\right)^2 \epsilon^{-2\theta}}. \tag{13}$$

The preceding formulae make it possible to calculate rigorously the transmission loss caused by any network whose image impedances and transfer constant are both known. In the symmetrical case, if $Z_I = Z_S = Z_R$, the transmission loss is determined simply by the value of the attenuation constant. In general, in the attenuation range of frequencies, the value of Θ of a wave filter is relatively large and the interaction factor is substantially unity. Consequently, the transmission loss caused by any filter in its attenuation range is dependent practically only upon the value of the attenuation constant and the reflection losses between Z_S and Z_{I_1}, Z_R and Z_{I_2}, and Z_S and Z_R, respectively. Throughout most of the transmission range of a filter, its image impedances may be made very closely equal to the terminating impedances so that the transmission loss caused by the filter in this range is dependent simply upon its attenuation constant. In the intervening range, between the attenuated and the non-attenuated bands, the transfer factor, the reflection factors and the interaction factor must all be taken into account.[7]

[7] Zobel, O. J., "Transmission Characteristics of Electric Wave-Filters," *Bell Sys. Tech. Jour.*, Oct., 1924.

Impedance and Propagation Characteristics of Non-Dissipative Filters. If the series and shunt impedances of the structures shown in Figs. 2 and 4 are pure reactances, as they would be in the case of a non-dissipative filter, the ratio of the quantity $Z_1/4Z_2$ must be either a positive or negative numeric. It has been shown by Campbell[8] and others that the attenuation constant is zero, and that the structure freely transmits at all frequencies at which the ratio $Z_1/4Z_2$ lies between 0 and -1. Therefore, by plotting values of the ratio $Z_1/4Z_2$ it is possible to determine the attenuation characteristic of any symmetrical structure as a function of frequency.

In the transmission range, the phase constant of the symmetrical structure shown in Fig. 2 or Fig. 4, is

$$B=2\sin^{-1}\sqrt{\frac{-Z_1}{4Z_2}}. \tag{14}$$

Hence, the expression for the image transfer constant of either of the symmetrical structures shown in Fig. 2 or Fig. 4 is

$$\theta=0+j\,2\sin^{-1}\sqrt{\frac{-Z_1}{4Z_2}}. \tag{15}$$

In the attenuation region, $Z_1/4Z_2$ may be either negative or positive. If $Z_1/4Z_2$ is negative and is greater in absolute magnitude than unity, the attenuation constant is

$$A=2\cosh^{-1}\sqrt{\frac{-Z_1}{4Z_2}} \tag{16}$$

and the phase constant, or the imaginary component of the image transfer constant, is

$$B=(2K-1)\pi \tag{17}$$

where K is any integer. Hence,

$$\theta=2\cosh^{-1}\sqrt{\frac{-Z_1}{4Z_2}}+j(2K-1)\pi. \tag{18}$$

From equation (8), when $Z_1/4Z_2$ is positive, the attenuation constant is

$$A=2\sinh^{-1}\sqrt{\frac{Z_1}{4Z_2}} \tag{19}$$

and the phase constant B is zero. Hence,

$$\theta=2\sinh^{-1}\sqrt{\frac{Z_1}{4Z_2}}+j0. \tag{20}$$

[8] Campbell, G. A., "Physical Theory of the Electric Wave-Filter," *Bell Sys. Tech. Jour.*, Nov., 1922.

As a result of equations (18) and (20), in the attenuation range, the phase constant of a non-dissipative symmetrical filter section is always zero or an odd multiple of $\pm\pi$.

The *cut-off frequencies*, by which are meant the divisional frequencies which separate the transmission bands from the attenuation bands, must always occur when $Z_1/4Z_2=0$ or when $Z_1/4Z_2=-1$, since, for the transmission bands, $Z_1/4Z_2$ must lie between 0 and -1.

The general formulae for the image impedances of the symmetrical networks shown in Figs. 2 and 4 are equations (6) and (7), respectively. From these equations, the image impedances are pure resistances in the transmission range of a non-dissipative structure. In the attenuation range, however, the image impedances are pure reactances; the mid-series image impedance is a reactance having the same sign as Z_1, while the mid-shunt image impedance is a reactance having the same sign as Z_2. In these attenuation bands, the image impedances (pure reactances) have positive or negative signs depending upon whether they are increasing or decreasing with frequency. The order of magnitude of the image impedances may be found from Table I.

TABLE I

If the Value of Z_1 is	And if the Value of $4Z_2$ is	Then the Mid-Series Image Impedance is	And the Mid-Shunt Image Impedance is
Zero	Zero	Zero	Zero
Zero	Finite	Zero †	Zero †
Zero	Infinite	Finite †	Finite †
Finite	Zero	Finite **	Zero **
Finite	Finite	Zero* or Finite	Infinite * or Finite
Finite	Infinite	Infinite †	Infinite †
Infinite	Zero	Infinite **	Zero **
Infinite	Finite	Infinite **	Finite **
Infinite	Infinite	Infinite	Infinite

* When both Z_1 and Z_2 are finite and $Z_1=-4Z_2$, the mid-series image impedance is zero and the mid-shunt image impedance is infinite.

† This condition gives a cut-off frequency.

** This condition results in infinite attenuation.

Types of Non-Dissipative Series-Shunt Sections Having Not More Than One Transmission Band or More Than One Attenuation Band. Since the series and shunt arms of a non-dissipative filter section may each be composed of any combination of pure reactances, it is possible to have an infinite number of types of filter sections. However, it is seldom desirable to employ filters having more than one transmission band or more than one attenuation band. Under these conditions, it is generally impracticable to employ more than four reactance elements in either of the arms of a section. Likewise, a total of six reactance elements in both the series and shunt arms is the maximum that can be economically employed.

Types of two-terminal reactance meshes having not more than four elements, are listed in Fig. 5. In Fig. 6, the corresponding frequency-

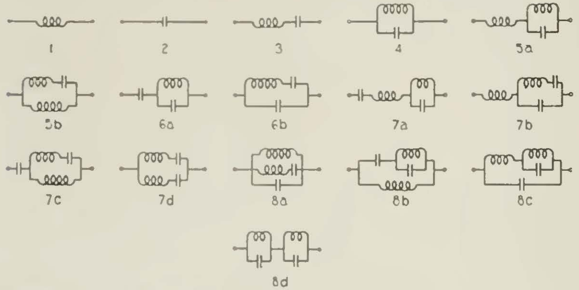


Fig. 5—Two-Terminal Reactance Meshes Containing Not More Than Four Elements

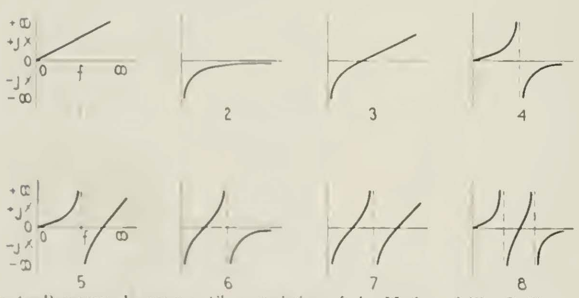


Fig. 6—Reactance-Frequency Characteristics, of the Meshes of Fig. 5, Shown in Symbolic Form

reactance characteristics are represented. Reactance characteristics Nos. 1 and 2 of Fig. 6 are reciprocal in nature, that is, their product is a constant, independent of frequency. Reactance characteristics Nos. 3 and 4 are similarly related if the frequencies of resonance and anti-resonance coincide. Similar relations exist between characteristics Nos. 5 and 6, and between characteristics Nos. 7 and 8. Two forms of reactance mesh in Fig. 5 (Nos. 5a and 5b) give the same reactance characteristic (No. 5 of Fig. 6) and are, therefore, by proper design, electrically equivalent. Characteristic No. 6 of Fig. 6 also corresponds to two reactance meshes of Fig. 5 (Nos. 6a and 6b) and the latter may, therefore, be considered equivalent. Likewise, reactance meshes 7a, 7b, 7c and 7d of Fig. 5 give characteristic No. 7 of Fig. 6 and are therefore potentially equivalent; also reactance

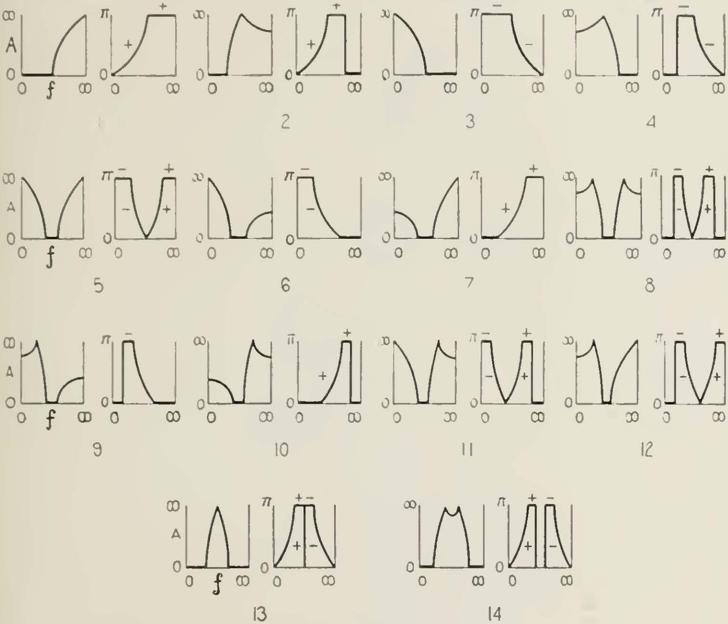


Fig. 7—Propagation Constant (Attenuation Constant and Phase Constant) Characteristics, Shown in Symbolic Form

meshes Nos. 8a, 8b, 8c and 8d of Fig. 5 are represented by reactance characteristic No. 8 of Fig. 6 and, consequently, may also be designed to be equivalent. The equivalence of the above reactance meshes has been discussed by Zobel [5] and will be subsequently treated at length. It is to be understood that, for the sake of brevity, in what follows, meshes Nos. 5, 6, 7 and 8 cover, respectively, all forms of the equivalent meshes: 5a and 5b; 6a and 6b; 7a, 7b, 7c and 7d; and

8a, 8b, 8c and 8d. Using these reactance combinations[9] for the series and shunt arms, there are only a relatively small number of types of filter structures. All of these types of filter structures are

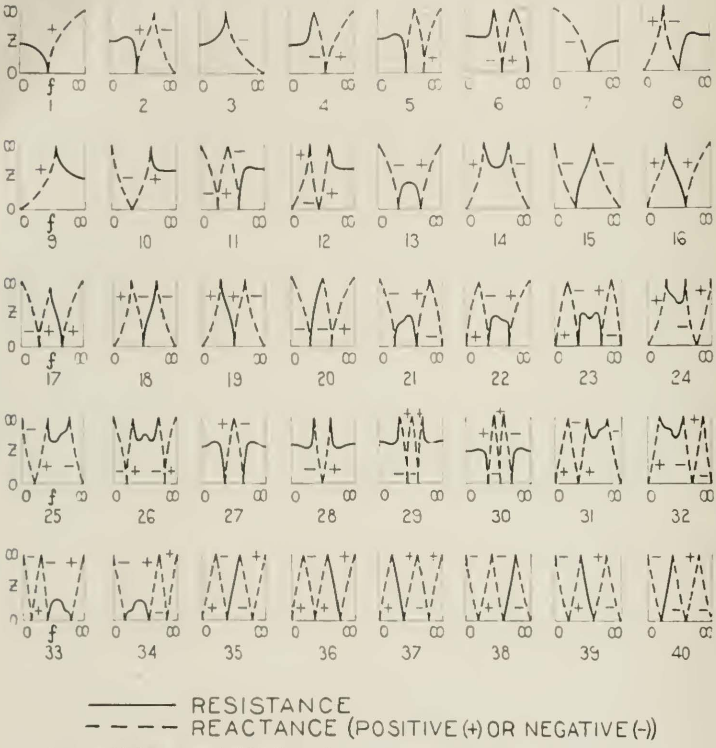


Fig. 8—Mid-Series and Mid-Shunt Image Impedance Characteristics, Shown in Symbolic Form

listed in Table II, and are called *low pass*, *high pass*, and *band pass* filters (having only one transmission band) and *band elimination*

[9] The general method of deriving the attenuation and phase characteristics of a section from the reactance frequency characteristics of its series and shunt arms is discussed by Zobel in Bibliography 13.

Tabulation of the Propagation and Impedance Characteristics of Series-Shunt Wave Filter Sections Which can be Formed by the Reactance Meshes Shown in Fig. 5

SERIES ARM

SHUNT ARM	1	2	3	4	5	6	7	8
1	No Pass Band	3-7-9	6-13-16	4-8-9	9-22-16	Band-and-High Pass	Double Band Pass	Band-and-High Pass
2	1-1-3	No Pass Band	7-13-15	2-2-3	Low-and-Band Pass	10-21-15	Double Band Pass	Low-and-Band Pass
3	2-1-4	4-7-11	9-13-17 10-13-20	14-27-28	2-5-4	4-11-10	9-33-17 10-34-20	14-30-28
4	7-16-14	6-15-11	5-13-14	9-18-11 10-19-14	12-22-14	11-21-14	Triple Band Pass	8-23-14
5	10-16-24	Band-and-High Pass	11-13-24	4-8-12	9-22-36 9-35-24 10-22-35 10-37-24	Double Band-and-High Pass	More Than Six Elements	More Than Six Elements
6	Low-and-Band Pass	9-15-25	12-13-25	2-2-6	Low-and-Double Band Pass	9-21-39 9-38-25 10-21-40 10-39-25	More Than Six Elements	More Than Six Elements
7	Low-and-Band Pass	Band-and-High Pass	8-13-26	14-27-29	More Than Six Elements	More Than Six Elements	More Than Six Elements	More Than Six Elements
8	Double Band Pass	Double Band Pass	Triple Band Pass	9-18-31 10-19-32	More Than Six Elements	More Than Six Elements	More Than Six Elements	More Than Six Elements

filters (having two pass bands and only one attenuation band). Their attenuation constant and phase constant characteristics, with respect to frequency, are shown symbolically in Fig. 7. The mid-series and mid-shunt image impedance characteristics with respect to frequency are shown in Fig. 8. In Table II, the figure at the head of each column indicates the reactance mesh in Fig. 5 which is used for Z_1 (series impedance) and the figure at the left of each row indicates the mesh in Fig. 5 which is used for Z_2 (shunt impedance). The figures in the squares of the table denote, reading from left to right, the propagation characteristics (attenuation and phase), the mid-series image impedance, and the mid-shunt image impedance, respectively, as shown in Figs. 7 and 8.

For example, the filter corresponding to the third column and to the fourth row (3 – 4) has a series arm composed of an inductance in series with a capacity as indicated by mesh 3 of Fig. 5, and has a shunt arm composed of an inductance in parallel with a capacity, as designated by mesh 4 of Fig. 5. The attenuation constant and phase constant characteristics of this filter are shown symbolically by diagram 5 of Fig. 7, while the mid-series and mid-shunt image impedances are indicated, respectively, by diagrams 13 and 14 of Fig. 8. The symbolic nature of the diagrams lies in the fact that the abscissae of each diagram cover the frequency range from zero to infinity, and the ordinates of Figs. 7 and 8 cover the attenuation constant and the impedances from zero to infinity. For example, the structure cited has an attenuation constant characteristic (diagram 5 of Fig. 7) composed of a transmission band lying between two attenuation bands, the attenuation constant being infinite in one of them at zero frequency, and in the other, at infinite frequency. The phase constant of this structure is $-\pi$ radians in the lower of the two attenuation bands, increases from $-\pi$ to $+\pi$ radians in the transmission band (passing through zero), and is $+\pi$ radians throughout the upper of the two attenuation bands. The mid-series image impedance (diagram 13 of Fig. 8) is a negative reactance in the lower of the two transmission bands, decreasing from infinity, at zero frequency, to zero at the lower cut-off frequency, is a pure resistance throughout the transmission band, and is a positive reactance, increasing from zero to infinity, in the upper of the two attenuation bands. The mid-shunt image impedance characteristic (diagram 14 of Fig. 8) is reciprocal in nature, for this structure, to the mid-series image impedance characteristic. This type of filter also possesses, in the general case, a double band pass attenuation characteristic and corresponding phase and impedance characteristics. A discussion of such characteristics is outside the scope of this paper even though many of the structures listed in Table II will show, if completely analyzed, multi-band characteristics. Where no specific characteristics are listed in Table II, no low pass, high pass, single band pass, or single band elimination characteristics are obtainable with a filter section limited to six different reactance elements.

In Table II, a large number of the structures have identically the same types of attenuation constant and phase constant characteristics. For example, six of the seven low pass filter sections have attenuation constant and phase constant characteristic No. 2 of Fig. 7. Likewise, six of the high pass structures have attenuation constant and phase constant characteristic No. 1. Also, in Table II, band pass groups are to be found having respectively, the following propagation characteristics common to each group: 6, 7, 8, 9, 10, 11 and 12. Finally, ten of the eleven band elimination structures listed have propagation constant characteristic No. 14.

Although six of the seven low pass wave filters have the same attenuation constant and phase constant characteristics, the various image impedance characteristics differentiate the structures among themselves. Similar differentiations exist in the high pass, band pass, and band elimination groups of structures. In each of the four types of filter sections however, all of those structures having the same series reactance meshes (that is, having the same series configuration of reactance elements) may be designed to have the same mid-series image impedance characteristic and, similarly, all of those structures within each type having the same shunt reactance meshes, or configuration of elements, may be designed to have the same mid-shunt image impedance characteristic.

In view of the fact that some of the structures listed in Table II have the same attenuation and phase constants but have different impedance characteristics, the question arises as to the relative virtues of the latter. Furthermore, since certain of the structures have the same mid-series or mid-shunt image impedances but have different propagation characteristics, it is possible to join together such structures and obtain a composite structure which has no internal reflection losses, that is, one whose total transfer constant is the sum of the various transfer constants of the individual sections. In order to minimize reflection and interaction losses in the transmission range, it is generally desirable to use, at the terminals of the filter, sections whose image impedances closely simulate those of the terminal impedances to which the filter is connected. The choice presented by filter structures having different impedance characteristics but the same propagation characteristic is, therefore, of advantage. In the attenuation range this is also true where impedance conditions are imposed at the terminals of the filter.

One class of structures which possess desirable image impedances and whose characteristics are readily determined from simpler structures is the so-called derived *m*-type.[5] The simplest forms of derived

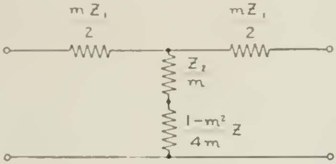


Fig. 9—Mid-Series Equivalent *m*-Type of Section

structures are shown in Figs. 9 and 10. The structure of Fig. 9 has the same mid-series image impedance as that shown in Fig. 2 and the value of this impedance is given by equation (6). The structure of Fig. 10 has the same mid-shunt image impedance as the π structure shown in Fig. 4 and the value of this impedance is given by

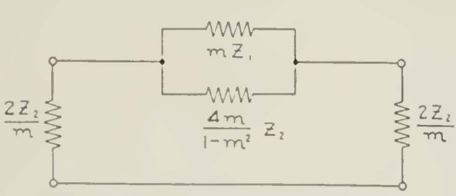


Fig. 10—Mid-Shunt Equivalent *m*-Type of Section

equation (7). On account of this identity of the respective mid-series and the mid-shunt image impedances in the two cases, the structures shown in Figs. 9 and 10 are called, respectively, the *mid-series equivalent derived m*-type and the *mid-shunt equivalent derived m*-type. The *T* and π structures of Figs. 2 and 4 are called, respectively, the ***prototypes*** of the derived *m*-structures of Figs. 9 and 10. In a series-shunt filter composed of sections of the *m*-type of Fig. 9 or Fig. 10, the ratio $(Z_1/4Z_2)_m$ of the series impedance to four times the shunt impedance is

$$\left(\frac{Z_1}{4Z_2}\right)_m = \frac{m^2\left(\dfrac{Z_1}{4Z_2}\right)}{1+(1-m^2)\left(\dfrac{Z_1}{4Z_2}\right)}. \tag{21}$$

From this expression, when $Z_1/4Z_2$ of the prototype is 0 or -1, the corresponding value of $(Z_1/4Z_2)_m$ for the derived m-type is also 0 or -1. Hence, the derived type has the same cut-off frequencies and

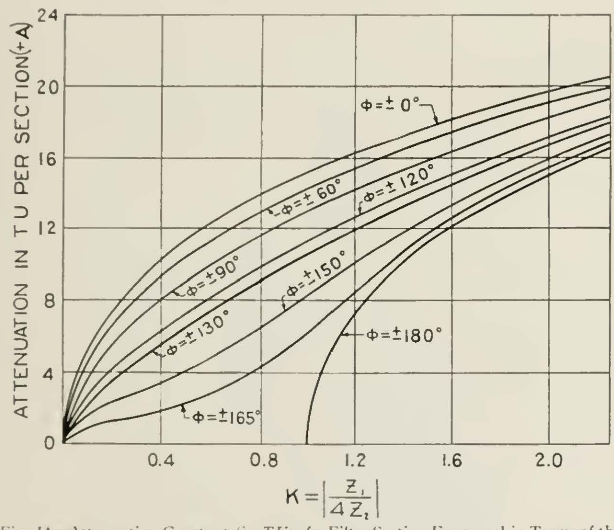


Fig. 11.—Attenuation Constant (in TU) of a Filter Section Expressed in Terms of the Ratio of Its Series Impedance to Four Times Its Shunt Impedance (i.e., $Z_1/4Z_2 = K/\Phi$)

therefore the same transmission and attenuation regions as its prototype.

Impedance and Propagation Characteristics of Dissipative Filters. It has been pointed out, in the case of non-dissipative structures, that the ratio $Z_1/4Z_2$ is either a positive or a negative numeric. If there is dissipation in the filter structure, that is, if the resistance associated with the reactance elements cannot be neglected, then the ratio

$Z_1/4Z_2$ will not, in general, be a numeric but a vector. However, the general formula (8), still holds true with dissipation. For determining the attenuation constant and phase constant of a dissipative structure it is convenient to use two formulae which may be derived from (8). These formulae are

$$A = \cosh^{-1}\left(K + \sqrt{(K-1)^2 + 4K\cos^2\frac{\phi}{2}}\right), \tag{22}$$

$$B = \cos^{-1}\left(-K + \sqrt{K^2 + 2K\cos\phi + 1}\right), \tag{23}$$

where

$$\frac{Z_1}{4Z_2} = \left|\frac{Z_1}{4Z_2}\right| \angle \pm\phi = K \angle \pm\phi.$$

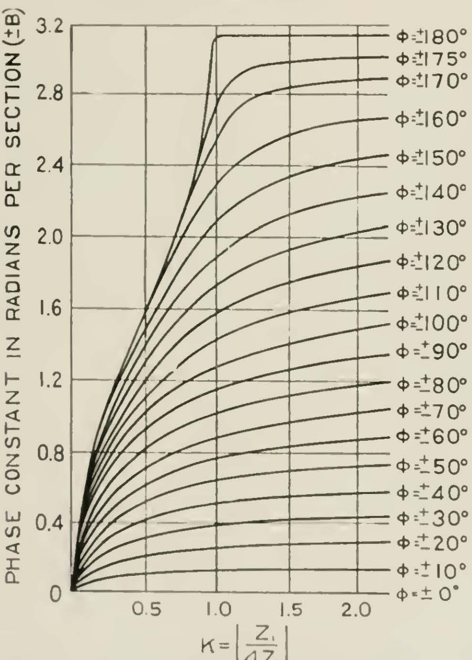


Fig. 12—Phase Constant of a Filter Section Expressed in Terms of the Ratio of Its Series Impedance to Four Times Its Shunt Impedance (i.e., $Z_1/4Z_2 = K\angle\pm\Phi$)

Formulae (22) and (23) are expressed in napiers and circular radians, respectively. They are represented in TU and in radians by families of curves such as are shown in Figs. 11 and 12.

A convenient ratio which expresses the dissipation in any reactance element is the absolute ratio, d, of its effective resistance to its reactance. In the case of a coil, $d = R/L\omega$ while in the case of a condenser $d = RC\omega$. The reciprocal ratio $Q = \frac{1}{d} = \frac{L\omega}{R} = \frac{1}{RC\omega}$ has also been widely used as a measure of dissipation in reactance elements. The ratio d or Q will not, in general, be constant over a wide frequency

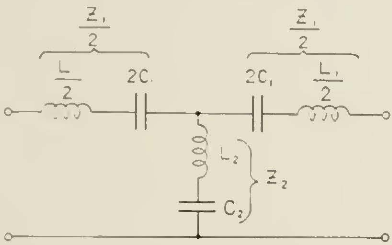


Fig. 13—Typical Band Pass Wave Filter Section (Mid-Series Termination)

range. If the value is known at an important frequency in the transmission range, it may ordinarily be regarded to hold for the rest of the transmission range. The effect of dissipation on the attenuation constant is most important in the transmission band, where the attenuation constant would be zero if there were no dissipation. Its effect is most pronounced in the neighborhood of the cut-off frequencies where the transmission bands merge into attenuation bands.

In the attenuation bands, the general effect of dissipation is negligible. It largely controls, however, the value of the attenuation constant at those frequencies at which infinite attenuation would occur if there were no dissipation. The effect of dissipation upon the phase constant is most pronounced in the neighborhood of the cut-off frequencies where resistance rounds off the abrupt changes in phase which would otherwise occur (see Fig. 12).

Characteristics of a Typical Filter. In order to illustrate specifically the principles employed in filter design, consider as an example the band pass structure 3–3 of Table II. This structure is illustrated in Fig. 13. It will be assumed that the dissipation in the coils cannot be neglected, but that the dissipation in the condensers is of negligible magnitude. If R_1 and R_2 are the effective resistances of the inductance elements L_1 and L_2, respectively, the series impedance, Z_1, of a series-shunt recurrent structure composed of sections of the type shown in Fig. 13 is

$$Z_1 = R_1 + j\left(\omega L_1 - \frac{1}{\omega C_1}\right). \tag{24}$$

The impedance of the shunt arm is

$$Z_2 = R_2 + j\left(\omega L_2 - \frac{1}{\omega C_2}\right). \tag{25}$$

In substituting for R_1 its value $L_1\omega d$ and for R_2 its value $L_2\omega d$, the ratio $Z_1/4Z_2$ becomes

$$\frac{Z_1}{4Z_2} = \frac{L_1}{4L_2} \cdot \frac{1 - jd - \dfrac{1}{\omega^2 L_1 C_1}}{1 - jd - \dfrac{1}{\omega^2 L_2 C_2}}. \tag{26}$$

Assuming d to be zero, the ratio $Z_1/4Z_2$ is

$$\frac{Z_1}{4Z_2} = \frac{C_2(\omega^2 L_1 C_1 - 1)}{4C_1(\omega^2 L_2 C_2 - 1)}. \tag{27}$$

Referring to Table II, the structure shown in Fig. 13 has two distinct attenuation and phase characteristics. These are, respectively, characteristics Nos. 9 and 10 of Fig. 7. These two sets of characteristics arise from the fact that the shunt arm may be resonant at a frequency less than, or greater than, the resonant frequency of the series arm. The two attenuation characteristics are inverse with respect to frequency. We shall, therefore, discuss only one of the two cases, namely, that in which the shunt arm resonates at a frequency greater than the resonant frequency of the series arm (that is, L_1C_1 is greater than L_2C_2). The frequency at which the shunt arm is resonant will be designated as f_∞, due to the fact that in a non-dissipative filter the attenuation constant is infinite at this point. In other words,

$$f_\infty = \frac{1}{2\pi\sqrt{L_2 C_2}}. \tag{28}$$

It is evident that the frequency at which Z_1 is resonant is a cut-off frequency since Z_1, and therefore $Z_1/4Z_2$, is zero at this point. An inspection of graphical curves[10] drawn for Z_1 and $4Z_2$, under the above

[10] For an illustration of the construction of such curves see Bibliography 12, Fig. 7, also Bibliography 13, Fig. 2.

conditions, will show that this is the lower of the two cut-off frequencies (f_1), that is

$$f_1 = \frac{1}{2\pi\sqrt{L_1C_1}}. \tag{29}$$

By equating $Z_1/4Z_2$ to -1 in equation (27) the upper cut-off frequency (f_2) is found to be

$$f_2 = \frac{1}{2\pi}\sqrt{\frac{C_2+4C_1}{C_1C_2(L_1+4L_2)}}. \tag{30}$$

For these explicit relations for f_1, f_2 and f_∞, equation (26) may be rewritten

$$\frac{Z_1}{4Z_2} = \left(\frac{f_1}{f_\infty}\right)^2 \frac{\left[\left(\frac{f_\infty}{f_2}\right)^2-1\right]\left[(1-jd)\left(\frac{f}{f_1}\right)^2-1\right]}{\left[1-\left(\frac{f_1}{f_2}\right)^2\right]\left[(1-jd)\left(\frac{f}{f_\infty}\right)^2-1\right]}. \tag{31}$$

When d is zero this equation becomes, for the non-dissipative case

$$\frac{Z_1}{4Z_2} = \frac{\left[1-\left(\frac{f_\infty}{f_2}\right)^2\right]\left[1-\left(\frac{f_1}{f}\right)^2\right]}{\left[1-\left(\frac{f_1}{f_2}\right)^2\right]\left[\left(\frac{f_\infty}{f}\right)^2-1\right]}. \tag{32}$$

From the preceding formulae and from the curves shown in Figs. 11 and 12, it is possible to read directly the attenuation constant and the phase constant for the structure shown in Fig. 13, at any frequency, provided the values of f_1, f_2 and f_∞ are known. The formulae for the dissipative case are of use mainly throughout the transmission bands and near the frequency f_∞. Elsewhere, the formulae for $Z_1/4Z_2$ for the non-dissipative structure may be employed without undue error. The preceding formulae have been derived in a direct manner, but may be obtained more simply by considering the structure of Fig. 13 to be a derived form of the structure 3−2 in Table II.

In order to minimize reflection loss effects, it is, as a rule, desirable to terminate a filter in an impedance equal to the image impedance of the filter at the mid-frequency,[11] (f_m) or at some other important frequency. From equation (6) and the values of Z_1 and Z_2, the mid-series image impedance (Z_o), at the mid-frequency in the non-dissipative case is

$$Z_o = \frac{1}{2}\left[\sqrt{\frac{L_1}{C_1}+4\frac{L_1}{C_2}} - \sqrt{\frac{L_1}{C_1}+4\frac{L_2}{C_1}}\right]. \tag{33}$$

[11] Defined as the geometric mean of the two cut-off frequencies f_1 and f_2; or $f_m = \sqrt{f_1 f_2}$.

From formulae (6), (29), (30), and (33) the mid-series image impedance at any frequency is

$$Z_I = Z_o\sqrt{1-\frac{\left(\frac{f}{f_m}-\frac{f_m}{f}\right)^2}{\left(\frac{f_2}{f_m}-\frac{f_1}{f_m}\right)^2}}. \tag{34}$$

An inspection of formula (34) indicates that the mid-series image impedance is symmetrical with respect to the mid-frequency, f_m.

In a similar way, the mid-shunt image impedance (Z_o') at the mid-frequency is

$$Z_o' = \sqrt{\frac{4L_1}{C_2\left(\frac{C_2}{C_1}+4\right)}} = \sqrt{\frac{4L_2}{C_1\left(\frac{L_1}{L_2}+4\right)}} \tag{35}$$

and the mid-shunt impedance, (Z_I'), at any frequency is

$$Z_I' = Z_o'\,\frac{1-\left(\frac{f}{f_r}\right)^2}{1-\left(\frac{f_m}{f_r}\right)^2}\sqrt{\frac{\frac{f_2}{f_1}-\left(\frac{f_m}{f}\right)^2}{\frac{f_2}{f_1}-\left(\frac{f}{f_m}\right)^2}} \tag{36}$$

It will be noted, that *if the values of the inductances and resistances of a filter are multiplied by any factor and if all the values of the capacities are divided by the same factor, the transmission loss-frequency characteristic is not changed*[12] (*neither are the cut-off frequencies, nor the frequencies of infinite attenuation*) *but the image impedances are multiplied by this factor.*

From the preceding formulae, explicit expressions may be derived for the values of L_1, C_1, L_2, and C_2. These expressions, which are given by Zobel,[5] in a slightly different form, are as follows:

$$L_1 = \frac{Z_o m}{\pi(f_2-f_1)}, \tag{37}$$

$$C_1 = \frac{f_2-f_1}{4\pi f_1^2 Z_o m}, \tag{38}$$

$$L_2 = \frac{Z_o}{\pi(f_2-f_1)}\cdot\frac{1-m^2}{4m}, \tag{39}$$

$$C_2 = \frac{(f_2-f_1)m}{\pi Z_o(f_2^2-f_1^2 m^2)}. \tag{40}$$

[12] Since the value of the transfer factor, e^{θ}, is dependent simply upon the ratio Z/Z_I, it is evident from equation (10) that *the transmission loss caused by the insertion of any network in a circuit is dependent simply upon impedance ratios.* Consequently, the above theorem is quite general and applies not only to filters but to any passive network.

where

$$m=\sqrt{1-\frac{\left(\frac{f_2}{f_1}\right)^2-1}{\left(\frac{f_\infty}{f_1}\right)^2-1}}. \tag{41}$$

As a numerical example of the determination of the constants of a filter section of the type under consideration, assume that the lower cut-off frequency, f_1, is 20,000 cycles, and that the upper cut-off frequency, f_2, is 25,000 cycles and that the frequency of infinite attenuation, f_∞, is 30,000 cycles. Assume, furthermore, that the value of the mid-series image impedance, Z_0, at the mid-frequency is 600 ohms. Then from formula (41), $m=.742$; hence from (37), $L_1=.0284$ henry; from (38), $C_1=.00224\times10^{-6}$ farad; from (39) $L_2=.00577$ henry and from (40) $C_2=.00486\times10^{-6}$ farad. Assuming $d=.01$, the value of $Z_1/4Z_2$ as given by formula (31) at f_m (22,360 cycles) is found to be $.305\,/176°.4$. Referring to formula (22), in which $K=.305$ and $\phi=176°.4$, or to the curves of Fig. 11, this value of $Z_1/4Z_2$ corresponds approximately to .041 napiers or .36 *TU*. Similarly, from equation (23), or from the curves of Fig. 12, this value of $Z_1/4Z_2$ gives 1.15 radians, or 67°, for the phase constant. At zero frequency, the value of $Z_1/4Z_2$ is, from equation (31), $.542\,/0°$, which corresponds to 1.36 napiers or to 11.8 *TU*. Likewise, at infinite frequency, the value of $Z_1/4Z_2$ is $1.23\,/0°$, which corresponds to an attenuation loss of 1.97 napiers or to 16.6 *TU*. From the curves of Fig. 12, the phase constant is zero both at zero and at infinite frequency.

Composite Wave Filters. It has previously been pointed out that certain groups of the structures listed in Table II have the same mid-series or mid-shunt image impedance characteristics but that the various structures in such a group may have different attenuation and phase constant characteristics.

If a filter is composed of any number of symmetrical or dissymmetrical sections, so joined together that the image impedances at the junction points of the sections are identical, the attenuation and phase constant characteristics of the composite structure so formed, are equal to the sum of the respective characteristics of the individual sections. Furthermore, the image impedances of the composite filter will be determined by the image impedances of the accessible ends of the terminating sections. The desirability of forming such composite filters arises from the fact that a better disposition of attenuation and phase can be obtained by employing, in one composite structure, a number of different types of the characteristics shown in Fig. 7.

The dissymmetrical networks ordinarily employed in composite structures are usually L type networks each of which may be regarded as one-half the corresponding symmetrical T or π network. Generalized forms of such networks are shown in Figs. 14A, B, and C. By joining two of these half-sections, such as are shown in Figs. 14B

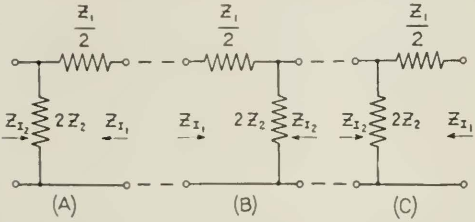


Fig. 14—Generalized Series-Shunt Structure Divided Into Successive Half-Sections (L-Type)

and C, we may form the full T section shown in Fig. 2. Similarly, by joining the two half-sections illustrated in Figs. 14A and B, the full π section of Fig. 1 results. The transfer constant, θ_{12}, of a half-section, such as is shown in Figs. 14A, B, or C, is one-half the transfer constant of the corresponding full section, that is,

$$\theta_{12} = \frac{\theta}{2} = \sinh^{-1}\sqrt{\frac{Z_1}{4Z_2}}. \tag{42}$$

Hence, the *attenuation constant and phase constant of a half-section* are, respectively, *one-half the attenuation constant and phase constant of a full section.* An important relationship between the half-section and the full section, which makes it convenient to use half-sections in composite wave filter structures, is that the image impedances, Z_{I_1} and Z_{I_2}, of any half-section are equal respectively to the mid-series and the mid-shunt image impedances of the corresponding full sections.

A typical example of the method of forming a composite low pass wave filter is given in Fig. 15, where three half-sections of different types and one full section are combined into a composite filter. The designations below the diagrams in Fig. 15A refer to the number of full sections and to the ratio f_r/f_c. In a practical filter, the various shunt condensers and series coils are combined as illustrated in Fig. 15B.

The composite nature of the attenuation characteristic of the filter of Fig. 15B is illustrated in Fig. 16, on a non-dissipative basis. In

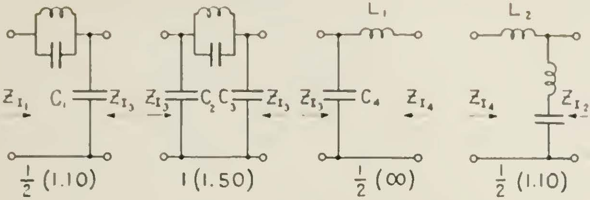


Fig. 15 A

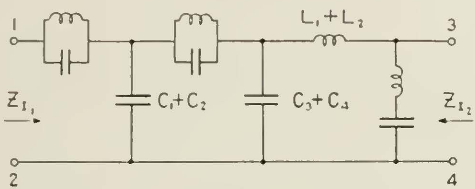


Fig. 15 B

Typical (Non-Dissipative) Composite Low Pass Wave Filter and Its Component Sections and Half-Sections

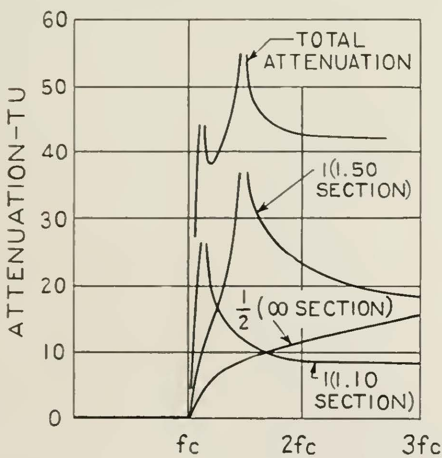


Fig. 16—Attenuation Characteristic of the Composite Low Pass Wave Filter of Fig. 15

Fig. 15B, the image impedance, Z_{I_1}, at the $1-2$ terminals has characteristic No. 2 of Fig. 8, while the image impedance, Z_{I_2}, at the $3-4$ terminals has characteristic No. 1 of Fig. 8.

Electrically Equivalent Networks. Reference has been made to the fact that *any passive network having one pair of input terminals and one pair of output terminals may be adequately represented, at any frequency, by an equivalent T or π network.* In general, this representation is a mathematical one and the arms of the T or π network cannot be represented, at all frequencies, by *physically realizable* impedances.

Furthermore, *any concealed network, containing no impressed electromotive forces, and having N accessible terminals is always capable of mathematical representation, at a single frequency, by a network having not more than $N(N-1)/2$ impedances, which impedances are determinable from the voltage and current conditions at the accessible terminals.* For networks having three or more terminals, this arbitrary mesh of impedances may possess a number of variant configurations. It is also true that the equivalence of the arbitrary mesh to the concealed network holds, at any single frequency, for any and all sets of external or terminal conditions, and that the magnitudes of the impedances of the arbitrary mesh are determinable, at will, on the assumption of the most convenient set of terminal conditions for each individual case. Familiar instances are the impedance equations derivable under various short-circuit and open-circuit conditions.

In specific cases, which are of particular interest, *one network may be shown to be capable of representation, as far as external circuit conditions are concerned, by another network which is physically realizable, and the latter may be substituted for the former, indiscriminately, in any circuit without consequent alteration, at any frequency, in the circuit conditions external to the interchanged networks.*

Equivalent meshes having two accessible terminals and employing respectively, three or four impedances in each mesh have been discussed by O. J. Zobel.[13] In filter design, two-terminal meshes are of importance only in those cases where the impedances are essentially reactances. Figs. 17A, B, C and D illustrate the physical configurations which reactance meshes employing not more than four elements may take. We are not generally interested in meshes having more than four elements for practical reasons which have previously been discussed. *Whenever any of the reactance meshes shown in Fig. 17 occur, we may, with proper design, substitute for it an equivalent mesh*

[13] See Appendix III of Bibliography 13.

of the associated type or types. Rigorous equivalence exists, *even with dissipation*, when the ratio of resistance to reactance, (d), is the same for all coils and the ratio of resistance to reactance (d') is the same for all condensers.

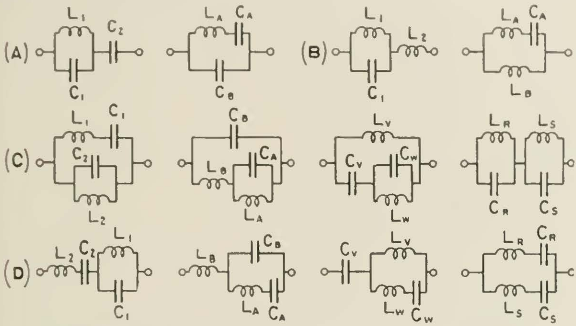


Fig. 17—Groups of Equivalent Two-Terminal Reactance Meshes

The relations which the equivalent meshes of Fig. 17 must observe are as follows:

17A

$$C_2=C_A+C_B,\quad C_1=\frac{C_B}{C_A}(C_A+C_B),\quad L_1=\frac{L_A}{\left(1+\frac{C_B}{C_A}\right)^2},\tag{43}$$

$$C_A=\frac{C_2^2}{C_1+C_2},\quad C_B=\frac{C_1C_2}{C_1+C_2},\quad L_A=L_1\left(1+\frac{C_1}{C_2}\right)^2,\tag{44}$$

17B

$$L_1=\frac{L_B^2}{L_A+L_B},\quad C_1=C_A\left(1+\frac{L_A}{L_B}\right)^2,\quad L_2=\frac{L_AL_B}{L_A+L_B},\tag{45}$$

$$L_A=\frac{L_2}{L_1}(L_1+L_2),\quad C_A=\frac{C_1}{\left(1+\frac{L_2}{L_1}\right)^2},\quad L_B=L_1+L_2,\tag{46}$$

17C

$$L_1=\frac{L_B}{L_A}(L_A+L_B)=L_W\left(1+\frac{C_W}{C_V}\right)^2=\frac{L_RL_S(L_R+L_S)(C_R+C_S)^2}{(L_RC_R-L_SC_S)^2}\tag{47}$$

$$L_2=L_A+L_B=L_V=L_R+L_S,\tag{48}$$

$$C_1 = \frac{C_A}{\left(1+\frac{L_B}{L_A}\right)^2} = \frac{C_V^2}{C_V + C_W} = \frac{(L_R C_R - L_S C_S)^2}{(L_R+L_S)^2 (C_R+C_S)}, \qquad (49)$$

$$C_2 = C_B = \frac{C_V C_W}{C_V + C_W} = \frac{C_R C_S}{C_R + C_S}. \qquad (50)$$

$$L_A = \frac{L_2^2}{L_1+L_2}, \quad L_B = \frac{L_1 L_2}{L_1+L_2}, \quad C_B = C_2, \qquad (51)$$

17C

$$C_A = C_1\left(1+\frac{L_1}{L_2}\right)^2, \quad L_W = \frac{L_1}{\left(1+\frac{C_2}{C_1}\right)^2}, \qquad (52)$$

$$L_V = L_2, \quad C_V = C_1 + C_2, \quad C_W = \frac{C_2}{C_1}(C_1+C_2), \qquad (53)$$

$$C_S = \frac{K + \sqrt{K^2 - 4L_2^2 C_1 C_2 K}}{2L_2^2 C_1}$$

where $K = (L_1C_1 + L_2C_1 + L_2C_2)^2 - 4L_1C_1L_2C_2$, (54)

$$C_R = \frac{C_S C_2}{C_S - C_2}, \quad L_S = \frac{L_1C_1 + L_2C_1 + L_2C_2 - L_2C_R}{C_S - C_R}, \quad L_R = L_2 - L_S. \qquad (55)$$

$$C_1 = \frac{C_B}{C_A}(C_A + C_B) = C_W\left(1+\frac{L_W}{L_V}\right)^2 = \frac{C_R C_S (C_R + C_S)(L_R + L_S)^2}{(L_R C_R - L_S C_S)^2}, \qquad (56)$$

$$C_2 = C_A + C_B = C_V = C_R + C_S. \qquad (57)$$

$$L_1 = \frac{L_A}{\left(1+\frac{C_B}{C_A}\right)^2} = \frac{L_V^2}{L_V + L_W} = \frac{(L_R C_R - L_S C_S)^2}{(C_R+C_S)^2(L_R+L_S)}, \qquad (58)$$

$$L_2 = L_B = \frac{L_V L_W}{L_V + L_W} = \frac{L_R L_S}{L_R + L_S}, \qquad (59)$$

$$C_A = \frac{C_2^2}{C_1 + C_2}, \quad C_B = \frac{C_1 C_2}{C_1 + C_2}, \quad L_B = L_2, \qquad (60)$$

17D

$$L_A = L_1\left(1+\frac{C_1}{C_2}\right)^2, \quad C_W = \frac{C_1}{\left(1+\frac{L_2}{L_1}\right)^2}, \qquad (61)$$

$$C_V = C_2, \quad L_V = L_1 + L_2, \quad L_W = \frac{L_2}{L_1}(L_1 + L_2), \qquad (62)$$

$$L_S = \frac{K + \sqrt{K^2 - 4L_1L_2C_2^2 K}}{2L_1C_2^2}$$

where $K = (L_1C_1 + L_1C_2 + L_2C_2)^2 - 4L_1C_1L_2C_2$, (63)

$$L_R = \frac{L_S L_2}{L_S - L_2}, \quad C_S = \frac{L_1C_1 + L_1C_2 + L_2C_2 - L_RC_2}{L_S - L_R}, \quad C_R = C_2 - C_S. \qquad (64)$$

For example, the two meshes in Fig. 17A will be equivalent if

$C_1 = .009$ mf.	$C_2 = .001$ mf.	$L_1 = .001$ h.
$C_B = .0009$ mf.	$C_A = .0001$ mf.	$L_A = .100$ h.

and the two meshes in Fig. 17B will be equivalent if

$L_1 = .002$ h.	$C_1 = .025$ mf.	$L_2 = .008$ h.
$L_A = .010$ h.	$C_A = .001$ mf.	$L_B = .010$ h.

Also, the four meshes of Fig. 17C will be equivalent if

$L_R = .001$ h.	$L_S = .002$ h.	$C_R = .001$ mf.	$C_S = .002$ mf.
$L_1 = .006$ h.	$L_2 = .003$ h.	$C_1 = .000333$ mf.	$C_2 = .000667$ mf.
$L_A = .001$ h.	$L_B = .002$ h.	$C_A = .003$ mf.	$C_B = .000667$ mf.
$L_V = .003$ h.	$L_W = .000667$ h.	$C_V = .001$ mf.	$C_W = .002$ mf.

and the four meshes of Fig. 17D will be equivalent if

$L_R = .001$ h.	$L_S = .001$ h.	$C_R = .001$ mf.	$C_S = .002$ mf.
$L_1 = .0000555$ h.	$L_2 = .0005$ h.	$C_1 = .024$ mf.	$C_2 = .003$ mf.
$L_A = .0045$ h.	$L_B = .0005$ h.	$C_A = .000333$ mf.	$C_B = .00267$ mf.
$L_V = .000555$ h.	$L_W = .005$ h.	$C_V = .003$ mf.	$C_W = .00024$ mf.

It is then evident that the following reactance meshes of Fig. 5 may be designed to be equivalent: 5a and 5b; 6a and 6b; 7a, 7b, 7c, and 7d; and 8a, 8b, 8c, and 8d. Hence, the following filter sections

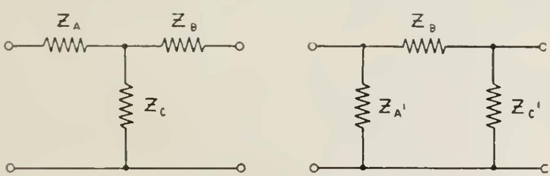


Fig. 18—Equivalent T and π Generalized Networks

referred to in Table II have, for the same impedance and propagation characteristics, a number of variant forms of physical configuration. 4−6, 6−2, 3−5, 6−4, 2−6, 5−3, 4−5, 1−5, 3−6, 5−4, 5−1, 4−8, 5−5, 6−6, 7−3, 6−3, 3−7, 4−7, 8−4 and 8−3.

Of the equivalent meshes having three accessible terminals the most common are the familiar T and π networks. The general relationships which must be observed for the equivalence of T or π networks are due to Kennelly [14] and for their generalized form, as illustrated in Fig. 18, are as follows:

$$Z_A = \frac{Z_A'Z_B'}{Z_A'+Z_B'+Z_C'}, \quad Z_B = \frac{Z_B'Z_C'}{Z_A'+Z_B'+Z_C'}, \quad Z_C = \frac{Z_A'Z_C'}{Z_A'+Z_B'+Z_C'}, \tag{65}$$

$$Z_A' = Z_A + Z_C + \frac{Z_AZ_C}{Z_B}, Z_B' = Z_A + Z_B + \frac{Z_AZ_B}{Z_C}, Z'_C = Z_B + Z_C + \frac{Z_BZ_C}{Z_A}. \tag{66}$$

We shall discuss here only two of the principal reactance meshes of the T and π form, namely, those employing solely inductances and

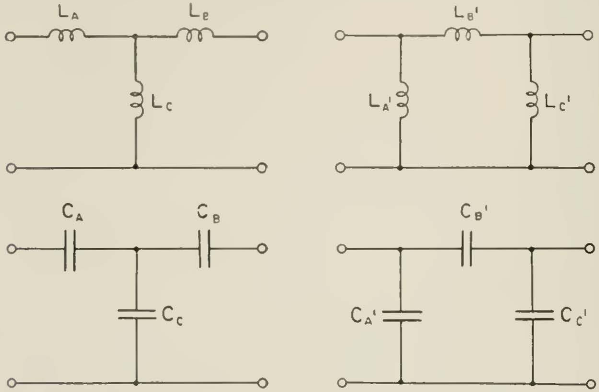


Fig. 19—Equivalent T and π Inductance Networks and Equivalent T and π Capacity Networks

solely capacities. It is to be understood that wherever an inductance or a capacity mesh of any of the following types occurs, its variant network may be substituted for it without change in the electrical characteristics of the circuit excluding those conditions within the mesh or its variant. Fig. 19 illustrates equivalent T and π networks of inductance and capacity.[15] The formulae relating the inductance and capacity meshes of Fig. 19 are as follows:

$$L_A = \frac{L_A'L_B'}{L_A'+L_B'+L_C'}, \quad L_B = \frac{L_B'L_C'}{L_A'+L_B'+L_C'}, \quad L_C = \frac{L_A'L_C'}{L_A'+L_B'+L_C'}, \tag{67}$$

[14] Kennelly, A. E., "The Equivalence of Triangles and Three-Pointed Stars in Conducting Networks," *Electrical World and Engineer*, New York, Vol. XXXIV, No. 12, pp. 413–414, Sept. 16, 1899. Also, "Application of Hyperbolic Functions to Electrical Engineering" (1911) (Appendix E).

[15] These meshes are rigorously equivalent, even when resistance is present if the ratio d is the same for all of the inductances and if the ratio d' is the same for all of the capacities.

$$L_A' = L_A + L_C + \frac{L_A L_C}{L_B},\quad L_B' = L_A + L_B + \frac{L_A L_B}{L_C},\quad L_C' = L_B + L_C + \frac{L_B L_C}{L_A}, \tag{68}$$

$$C_A' = \frac{C_A C_C}{C_A + C_B + C_C},\quad C_B' = \frac{C_A C_B}{C_A + C_B + C_C},\quad C_C' = \frac{C_B C_C}{C_A + C_B + C_C}, \tag{69}$$

$$C_A = C_A' + C_B' + \frac{C_A' C_B'}{C_C'},\qquad C_B = C_B' + C_C' + \frac{C_B' C_C'}{C_A'},$$

$$C_C = C_A' + C_C' + \frac{C_A' C_C'}{C_B'}. \tag{70}$$

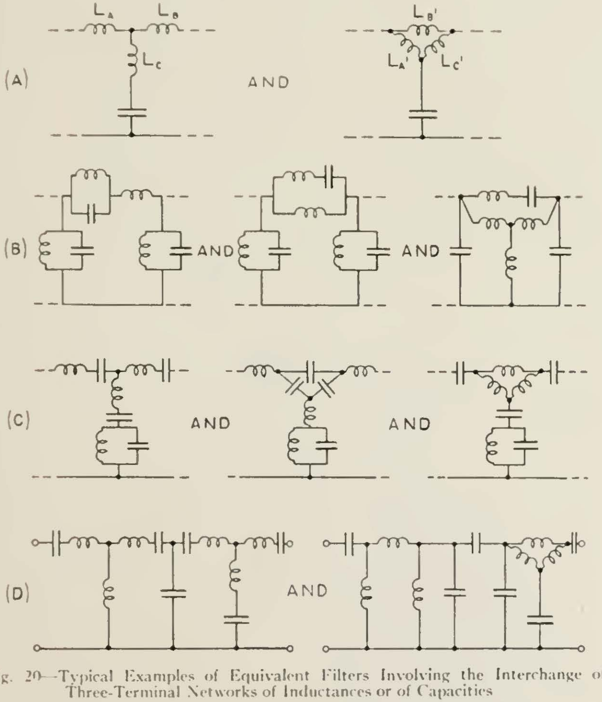


Fig. 20—Typical Examples of Equivalent Filters Involving the Interchange of Three-Terminal Networks of Inductances or of Capacities

A few examples of the variant filter structures which may arise, due to the existence of equivalent three terminal meshes of capacity and inductance, are illustrated in Fig. 20, in which Figs. 20A, B, and C represent either individual sections or portions of composite filters and Fig. 20D represents a composite filter. When equivalent reactance meshes occur entirely within a filter or within a section of a filter, the filter or the section will have the same cut-off frequencies and frequencies of infinite attenuation and the same attenuation, phase, and image impedance characteristics, whichever equivalent

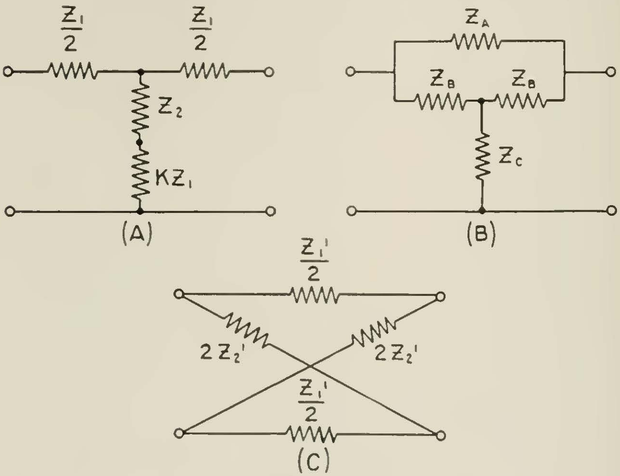


Fig. 21—Generalized Forms of Equivalent Series-Shunt, Bridged-*T*, and Lattice Type Filter Structures

form of mesh is substituted for an existing mesh. When equivalent meshes are interchanged in either recurrent or composite filters the substitution is generally made after the series-shunt structure is designed and after it has been found that the substitution will effect economies. The three terminal meshes referred to occur, in general, in unbalanced filter structures. For balanced filter circuits, corresponding meshes will be found for each of the equivalent networks by the process of dividing equally the series impedance between the two series lines of the filter.

While the discussion in this paper is based principally on the series-shunt structure there are two other important types of structures which will be mentioned. These are the so-called *lattice* [5] type structure and the *bridged-T* type structure. Typical series-shunt, bridged-T, and lattice type structures are illustrated in Fig. 21A, B and C, respectively. The three circuits shown are electrically equivalent, except for balance between the series arms, if the following relations hold:

$$Z_A = \left(1 + \frac{1}{4K}\right) Z_1, \qquad Z_B = (\tfrac{1}{2} + 2K) Z_1, \qquad Z_C = Z_2. \tag{71}$$

$$Z_1' = Z_1, \qquad Z_2' = (\tfrac{1}{4} + K) Z_1 + Z_2. \tag{72}$$

In the previous discussion of equivalent networks no reference has been made to networks containing mutual inductance, many of which are of particular interest and importance. These will be now discussed in detail.

PART II

Wave Filters Using Mutual Inductance

Before considering the equivalent meshes which may be formed by the use of mutual inductance between pairs of coils, and the types of wave filters which may be obtained by the use of these equivalent meshes, it will be necessary to define certain general terms.

The *self impedance* between any two terminals of an electrical network is the vector ratio of an applied e.m.f. to the resultant current entering the network when all other accessible terminals are free from external connections.

The *mutual impedance* of any network, having one pair of input terminals and one pair of output terminals, is the *vector ratio* of the e.m.f. produced at the output terminals of the network, on open circuit, to the current flowing into the network at the input terminals. Since mutual impedance is a vector ratio, it may have either of two signs, depending on the assumed directions of the input current and the output voltage. The sign of the mutual impedance is, in general, identified by its effect in increasing or decreasing the vector impedance of the meshes in which it exists. It is usually convenient, in this case, to consider either a simple series or a simple parallel mesh of two self impedances between which the mutual impedance acts. For the purpose of determining the sign of the mutual impedance, we shall confine our discussion to a simple series combination. Consequently, the mutual impedance will be called either *series aiding* or *series opposing*.

When a mutual impedance, Z_M, acts between two self impedances Z_1 and Z_2, (Fig. 22) connected in series in such a way as to *increase vectorially* the impedance of the combination, it is called a *series aiding* *mutual impedance.* Similarly, when a mutual impedance acts in such a way as to *decrease vectorially* the impedance of such a combination,

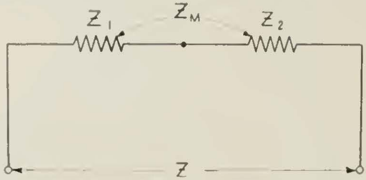


Fig. 22—Mutual Impedance Acting Between Two Self Impedances Connected in Series

it is called a *series opposing mutual impedance.* For example, if the total impedance, Z, of the combination shown in Fig. 22 is

$$Z = \frac{V}{I} = \frac{(IZ_1 + IZ_M) + (IZ_2 + IZ_M)}{I} = Z_1 + Z_2 + 2Z_M \qquad (73)$$

the mutual impedance is series aiding. On the other hand, if the total impedance, Z, of the combination is

$$Z = \frac{V}{I} = \frac{(IZ_1 + IZ_M) + (IZ_2 + IZ_M)}{I} = Z_1 + Z_2 - 2Z_M \qquad (74)$$

the mutual impedance is series opposing.

Transformer Representation. If, in Fig. 22, Z_1 represents the self impedance of one winding of a transformer and Z_2 the self impedance

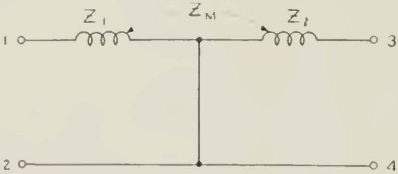


Fig. 23—*T* Network Containing Two Self Impedances, Having Mutual Impedance Between Them

of its other winding, the series impedance of the two windings (between terminals 1 and 3 in Fig. 23), as given by equations (73) and (74), will determine whether the mutual impedance, Z_M, is series aiding or series opposing.

The mutual impedance between the two windings may be represented by an equivalent network of self impedances connected as shown in Fig. 24. The four-terminal network illustrated in Fig. 24 may have various configurations. The equivalent T form is shown in Fig. 25. In view of the equivalence illustrated in Fig. 25,

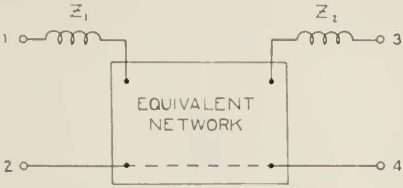


Fig. 24 Equivalent Network Representation of the Structure Shown in Fig. 23

the two-winding transformer of Fig. 23 may itself be completely represented by a single T network as indicated in Fig. 26. The theory of the equivalent T network representation of a transformer has been

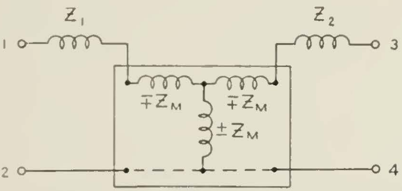


Fig. 25 T Network Representation of the Structure of Fig. 24

discussed by G. A. Campbell,[1] W. L. Casper[6] and others. In general, the self and mutual impedances of a transformer will be complex quantities. The arms of its equivalent T network will contain resist-

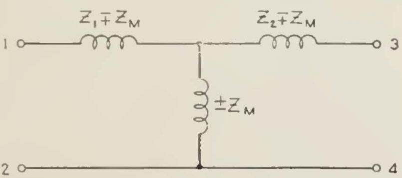


Fig. 26 T Network of Self Impedances Equivalent to the Structure of Fig. 23

ance and inductance components which may be either positive or negative. However, in the case of a transformer having no dissipation, (i.e., no *d-c.* resistance, no eddy current and no hysteresis losses) the arms of its equivalent T network are composed simply of positive or negative inductances. Of the three inductances involved, at least two of them must be positive while the third may be either positive or negative.

From Fig. 25, it is evident that two windings or coils, together with their mutual impedance, may be represented by an equivalent network which affords a transfer of energy from one winding to the other. This equivalent network may, with limitations, contain positive or negative inductances.

While the two-winding transformer of Fig. 23 has been represented by an equivalent T network in Fig. 26, the equivalent network may alternatively be of π form (Fig. 27) instead of T form, through the

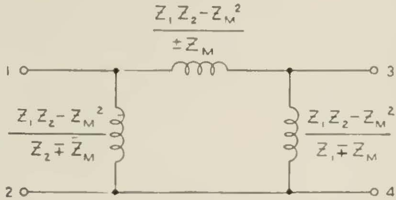


Fig. 27—π Network of Self Impedances Equivalent to the Structure of Fig. 23

general relationships for T or π networks previously stated. When no dissipation exists in the transformer, either equivalent network will have at least two positive inductances while the third inductance may be either positive or negative.

From the principles previously outlined in Part I, for the equivalence of certain electrical meshes and for their substitution for one another in any circuit, it is obvious that when two coils, with mutual impedance between them, exist in a circuit, in the manner shown in Fig. 23, either of the meshes shown in Fig. 26 or 27 may be substituted for them or vice versa. The representation of the mutual impedance, Z_M, by an equivalent network (Fig. 25) makes it possible to represent the transformer of Fig. 23 by a T or π network containing only self impedances. This affords a great simplification in the analysis of filter circuits containing pairs of coils having mutual impedance between them in that it permits such circuits to be reduced to an equivalent series-shunt (or lattice or bridged-T) type structure. Consequently, the methods of design which have been built up for the series-shunt and kindred type structures may be directly applied to the solution of circuits containing such pairs of coils.

Two-Terminal Equivalent Meshes. A list of equivalent two-terminal reactance meshes, due to Zobel, has been given in Fig. 17. All of the meshes in Figs. 17B, C and D contain two inductance elements. Mutual inductance may exist between any two inductive elements without changing fundamentally the nature of the reactance meshes. This means that when mutual inductance exists between two coils in

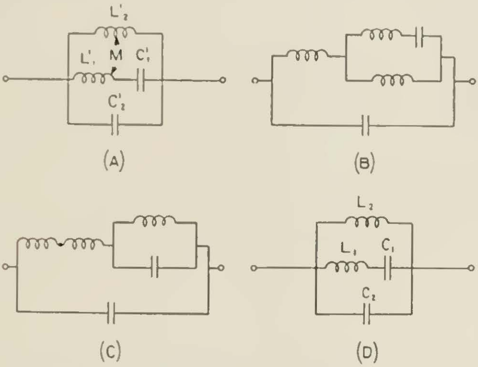


Fig. 28—Equivalent Two-Terminal Reactance Networks, Only One of Which Contains Mutual Inductance

any of these meshes, the mesh may be designed to be electrically equivalent to, and consequently can be substituted for, a corresponding mesh of the same type having no mutual inductance.

For example, consider the mesh shown in Fig. 28A which is potentially equivalent to the first reactance mesh of Fig. 17C and, consequently, to the other three reactance meshes of the same figure. The inductance elements L_1' and L_2', together with the mutual inductance M acting between them, may be represented by an equivalent T network, as previously stated. The reactance mesh formed by L_1', L_2', and M, together with its equivalent T and π forms, is shown in Fig. 29. By means of the relations given in Figs. 29A and B, it is possible to derive, from the structure of Fig. 28A, the equivalent structure shown in Fig. 28B. Likewise, from formulae (45) and (46) for the equivalence of the two structures of Fig. 17B, the mesh of Fig. 28C can be obtained from that of Fig. 28B. Furthermore, if the two inductances shown in series in Fig. 28C are merged, it is again possible, by means of the conversion formulae for the two meshes of

Fig. 17B, to determine the constants of the mesh shown in Fig. 28D from the known values of the constants of the structure of Fig. 28C.

The relations which must exist if the structure of Fig. 28D is to be equivalent to the structure shown in Fig. 28A, or vice versa, are given by the following relations

$$C_2 = C_2', \ L_1 = \frac{L_2'(L_1'L_2' - M^2)}{(L_2' \pm M)^2}, \tag{75}$$

$$L_2 = L_2', \ C_1 = C_1' \left(\frac{L_2' \pm M}{L_2'} \right)^2. \tag{76}$$

The upper and lower of the alternative signs, in the preceding equations, correspond respectively to series aiding and opposing connections. The equivalence of these four-element meshes makes it possible

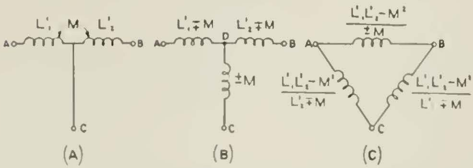


Fig. 29—Equivalent Three-Terminal Inductance Networks

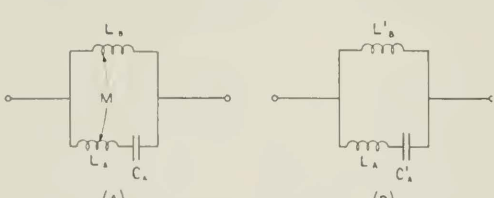


Fig. 30—Equivalent Two-Terminal Reactance Networks, Only One of Which Contains Mutual Inductance

to derive at once, the relations which must exist between certain equivalent three-element meshes involving mutual inductance. For example, if the capacity C_2' of Fig. 28A is zero, the mesh reduces to the three-element mesh of Fig. 30A and the formulae given above are then applicable for the equivalence of the structures of Figs. 30A and B.

In the same way that the meshes illustrated in Fig. 28 were shown to be potentially equivalent to each other, it is possible to prove that the meshes of Fig. 31 are potentially equivalent. The equivalence of the mesh shown in Fig. 31B to that of Fig. 31A is satisfied by the relations given in Figs. 29A and B. The equivalence of the mesh of Fig. 31C to that of Fig. 31B is governed by the equations (56 to 64) for the equivalence of the first and last structures of Fig. 17D. Finally, the equivalence of the mesh of Fig. 31D to that of Fig. 31C is controlled by the relations for the equivalence of the first two structures of Fig. 17D.

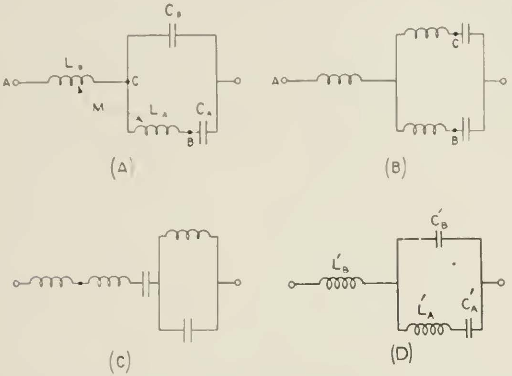


Fig. 31 Equivalent Two-Terminal Reactance Networks, Only One of Which Contains Mutual Inductance

The formulae relating the constants of the structure shown in Fig. 31D to the corresponding constants of the structure shown in Fig. 31A are as follows:

$$L_A' = L_1\left(1+\frac{C_1}{C_2}\right)^2,\ L_B' = L_2,\ C_A' = \frac{C_2^2}{C_1+C_2},\ C_B = \frac{C_1C_2}{C_1+C_2}, \qquad (77)$$

in which—

$$C_1 = \frac{C_AC_B(C_A+C_B)L_A^2}{[C_A(L_A \pm M) \pm MC_B]^2},\ C_2 = C_A + C_B, \qquad (78)$$

and

$$L_1 = \frac{[C_A(L_A \pm M) \pm MC_B]^2}{(C_A+C_B)^2L_A},\ L_2 = \frac{L_AL_B - M^2}{L_A}. \qquad (79)$$

The upper and lower of the alternative signs, in the preceding equations correspond, respectively, to series aiding and opposing connections.

The equivalence of these four-terminal meshes makes it possible to derive the relations which must exist for corresponding equivalent three-element meshes, with and without mutual inductance. For example, if in Fig. 31A, the capacity C_A is of infinite value, the mesh reduces to that shown in Fig. 32A and the formulae given above are applicable for the equivalence of the meshes of Figs. 32A and B.

The remaining meshes of Figs. 17C and D have similar potential equivalence to meshes of the same fundamental type but having mutual inductance between the respective pairs of coils.

Three-Terminal Equivalent Meshes. Three terminal meshes containing mutual inductance will now be discussed. It has been shown

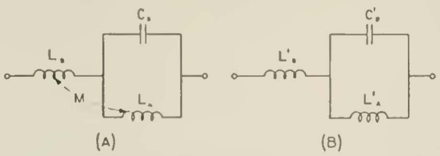


Fig. 32—Equivalent Two-Terminal Reactance Networks, Only One of Which Contains Mutual Inductance

that two coils, with mutual inductance between them (Fig. 29A), are equivalent to certain T and π structures containing only tangible inductances (Figs. 29B and C). Referring to Fig. 29B, it is seen that two coils, with series opposing mutual inductance between them (corresponding to the upper alternative signs in Fig. 29B), are equivalent to a T network having three positive inductance arms, provided the mutual inductance M is less than L_1' and L_2'. The values of these arms are respectively, $L_1'-M$, $L_2'-M$, and M. If M is larger than L_1', one arm of the equivalent T network is a negative inductance while the other two arms are positive inductances. Similarly, if M is larger than L_2', a different arm of the T network will be a negative inductance while the two remaining arms will be positive inductances. It is physically impossible for the value of M to be greater than both L_1' and L_2'. Hence, it is impossible for more than one arm of the T network, shown in Fig. 29B, to be a negative inductance.

When two coils have series aiding mutual inductance between them (the lower of the alternative signs in Fig. 29B) they are equivalent to a T network in which two of the arms consist of positive inductances viz., $L_1'+M$ and $L_2'+M$, while the third arm consists of a negative inductance of the value $-M$.

Whenever, in an equivalent T network, one of the arms is a positive (or negative) inductance, a corresponding arm of the π network will also be a positive (or negative) inductance. Consequently, as in the case of the equivalent T network, the equivalent π network shown in Fig. 29C may consist of three positive inductances or two positive inductances and one negative inductance, depending upon the sign and magnitude of M.

It is interesting to note that, in Fig. 29B, point D is in reality a concealed terminal, i.e., it cannot be regarded as physically accessible. There are, therefore, only three accessible terminals to the equivalent

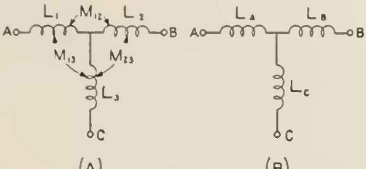


Fig. 33—Equivalent T Networks of Inductance

T network. In the π network shown in Fig. 29C there is no such concealed point. There are, however, as in the preceding case, three accessible terminals A, B and C.

When the mutual inductance, M, is equal to either one of the self inductances, L_1' (or L_2'), and the windings are connected in series opposing, the equivalent T and π networks of the transformer coalesce to the same L type network. For example, if $L_1'=M$ in Fig. 29A both the T and the π networks of Figs. 29B and C resolve into an L network whose vertical arm has the value M and whose horizontal arm is $L_2'-M$.

A problem of practical importance is the equivalence of T and π meshes, containing three coils with mutual inductance between all of the elements, to similar T and π meshes containing no mutual inductance. The T networks of Fig. 33 are potentially equivalent. The formulae governing their equivalence are

$$L_A=L_1+M_{12}+M_{13}-M_{23}, \tag{80}$$

$$L_B=L_2+M_{12}-M_{13}+M_{23}, \tag{81}$$

$$L_C=L_3-M_{12}+M_{13}+M_{23}. \tag{82}$$

In the above formulae, the signs correspond to the case of a series aiding mutual inductance between all the pairs of coils. When the mutual inductance between any two coils changes sign, the signs accompanying that mutual inductance in the above formulae are reversed.

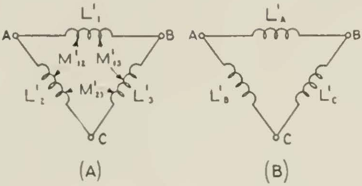


Fig. 34—Equivalent π Networks of Inductance

Similarly, the π networks of Fig. 34 are also potentially equivalent. The formulae governing their equivalence are

$$L_A' = \frac{L_xL_y + L_xL_z + L_yL_z}{L_y}, \tag{83}$$

$$L_B' = \frac{L_xL_y + L_xL_z + L_yL_z}{L_z}, \tag{84}$$

$$L_C' = \frac{L_xL_y + L_xL_z + L_yL_z}{L_x}, \tag{85}$$

in which

$$L_x = \frac{L_A''L_B''}{L_A'' + L_B'' + L_C''} \mp M_{12}', \tag{86}$$

$$L_y = \frac{L_B''L_C''}{L_A'' + L_B'' + L_C''} \mp M_{23}', \tag{87}$$

$$L_z = \frac{L_A''L_C''}{L_A'' + L_B'' + L_C''} \mp M_{13}', \tag{88}$$

where

$$L_A'' = L_1' \pm M_{12}' \pm M_{13}', \tag{89}$$

$$L_B'' = L_2' \pm M_{12}' \pm M_{23}', \tag{90}$$

$$L_C'' = L_3' \pm M_{13}' \pm M_{23}'. \tag{91}$$

As in the preceding case, the upper of the two signs occurs with the series aiding mutual inductance between all the pairs of coils. When the mutual inductance between any two coils changes sign, the signs accompanying that mutual inductance in the above formulae are reversed.

At least two of the three inductances (in Fig. 33B or in Fig. 34B) will always be positive in sign while the third inductance may be either positive or negative. Consequently, three coils having mutual inductance between each of them and having only three accessible terminals offer no greater possibilities than do two coils having mutual inductance between them and having three terminals. In both cases the structure is equivalent to a T or π mesh composed of three self

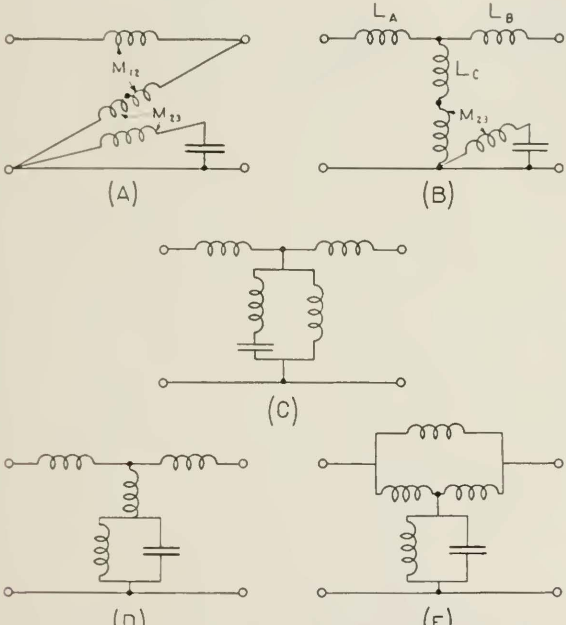


Fig. 35—Equivalent Filter Sections, With and Without Mutual Inductance

inductances, at least two of which must be positive. With specific relations between the various self and mutual inductances, it is possible for the three coils with mutual inductance between each of them to be equivalent (as in the case of two coils with mutual inductance) simply to an L network composed of two positive self inductances.

Since either two or three coils with mutual inductance between them are, in general, equivalent, at all frequencies, to a T or π network composed of three self inductances, it is possible to substitute the one type of mesh for the other in any kind of a circuit without affecting the currents or voltages external to the meshes involved. This substitution is always physically possible provided none of the arms of the equivalent T or π networks is a negative inductance.

The structures shown in Fig. 35 are illustrative of the power of equivalent networks as tools for the solution of filter structures containing mutual inductance. The equivalence of the structure shown in Fig. 35B to that of Fig. 35A is evident from the equivalence of two coils (Fig. 29) with mutual inductance (M_{12}) between them to three inductances, L_A, L_B and L_C without mutual inductance. Likewise,

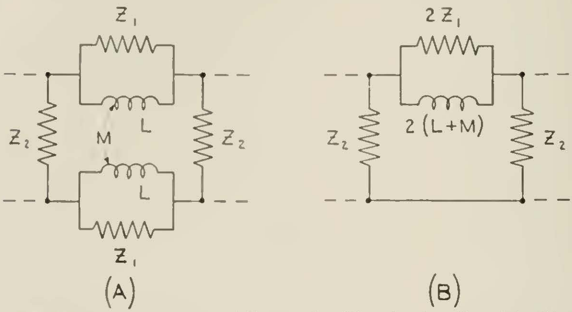


Fig. 36—Balanced and Unbalanced Forms of a Filter Section, Containing Mutual Inductance

the equivalence of the structure shown in Fig. 35C to that of Fig. 35B is obtainable by successive mesh substitutions. The equivalence of the structures shown in Fig. 35D and E to that of Fig. 35C are also obtainable from equivalences previously referred to. If the propagation and impedance characteristics of either of the structures of Fig. 35C or D are known, then the other structures shown in Fig. 35 will have the same characteristics. Furthermore, if the values of the constants of any one of these structures are known, the constants of any of the other structures are readily obtainable by means of transformation formulae.

In a large number of wave filters, the structures are unbalanced; that is, all of the series impedances are placed in one of the two line wires while the remaining wire is a short circuit. Ordinarily, the object in using such an unbalanced structure is to minimize the number of elements required in the series arms. It should be noted, however, (Fig. 36) that in case an inductance element enters into both series arms, it can be replaced, in symmetrical structures, by two equal windings of a single coil having mutual inductance between them and of such value that the series aiding inductance of these two coils is equal to the total inductance required in the corresponding unbalanced structure. For example, the structures shown in Figs. 36A and B are electrically equivalent to each other, that is, they have the same image impedance and transfer constant.

Types of Sections Obtainable Whose Equivalent Series-Shunt Sections Contain No Negative Inductances. It has previously been stated that an infinite number of types of series-shunt filter sections may be had, if no limitations are placed on the complexity of their reactance arms. It has also been stated, however, that for filters employing only one transmission or one attenuation band, the maximum number of elements which can ordinarily be used economically per section is six. A similar limitation exists when mutual inductance is employed, in that sections can seldom be economically used whose prototype structures contain more than six reactance elements.

Inasmuch as by the equivalences which have been discussed, many varient forms of a section may exist, which forms are reducible to the same series-shunt prototype, an effort only to list and discuss the prototype sections will be made. The prototype to which any given section then reduces will readily be found by the application of the foregoing principles. A few examples will later serve to make this clear.

In considering the prototype sections which exist when mutual inductance is present in a filter section, we shall first list the reactance meshes of which mutual inductance may form a part. Referring to Fig. 5, an inspection of the equivalences so far discussed will show that the following meshes may be partly or wholly composed of mutual inductance:

1, 3, 4, 5 (*a* and *b*), 7 (*a* and *b*), and 8 (*a* and *b*).

Consequently, a large number of the sections listed in Table II and formed from the reactance meshes of Fig. 5 may represent *not only actual sections containing no mutual inductance, but also equivalent prototypes of sections containing mutual inductance.* Sections containing mutual inductance within only the series arm or the shunt arm, respectively, are not included in this discussion since such arms may be readily reduced to equivalent arms, without mutual inductances, by the substitution of equivalent two-terminal meshes. The prototypes which are under discussion are listed below:

Low pass	*High pass*
1−3, 5−3	4−1, 4−5

Band pass

3 − 1, 1 − 1, 3 − 3, 1 − 1, 1 − 5, 5 − 1, 3 − 7, 3 − 5, 8 − 1, 1 − 8, 5 − 4, 5 − 5, and 7 − 3.

Sections corresponding to the equivalent series-shunt prototypes listed will have the same impedance and propagation characteristics as the prototype, and may be used indiscriminately in place of the prototype. Consequently, when a section has been reduced to any of the above prototypes, its various characteristics may be found from Table II and Figs. 7 and 8.

As an example of structures which have mutual inductance and which are equivalent to structures listed above, consider the section

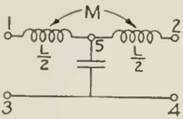


Fig. 37—Low Pass Filter Section Containing Two Coils, Having Mutual Inductance Acting Between Them, and a Condenser Shunted From Their Junction Point

shown in Fig. 37. This section contains two coils having mutual inductance, and a condenser shunted from their junction point. The three-terminal mesh formed by the two coils $L/2$ and $L/2$, together with their series opposing mutual inductance M, may be represented, as in Fig. 29B, by its equivalent T mesh. The resulting equivalent section is that shown in Fig. 38. The structure of Fig. 38, having a series reactance mesh corresponding to No. 1 of Fig. 5, and a shunt

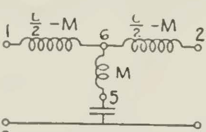


Fig. 38—Filter Section Containing No Mutual Inductance, Equivalent to the Section of Fig. 37

reactance mesh corresponding to No. 3 of Fig. 5 is that listed as 1 − 3 in Table II and in the above list. Consequently, it has propagation characteristic No. 2 of Fig. 7, and mid-series image impedance characteristic No. 1 of Fig. 8. The section of Fig. 37 may, consequently, be joined at either end to any structure having a mid-series image impedance characteristic such as that designated as characteristic No. 1 of Fig. 8. The section of Fig. 37 is not capable of mid-shunt termination since point 6 of Fig. 38 is not physically accessible.

Similarly, the section shown in Fig. 39 is equivalent to the series-shunt structure of Fig. 10. If the transformer mesh in Fig. 39, formed by $2L_2$, M and $2L_2$ be replaced by its equivalent π mesh, assuming series opposing windings the structure of Fig. 40 results.

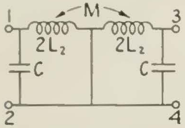


Fig. 39 Band Pass Filter Section Containing Mutual Inductance

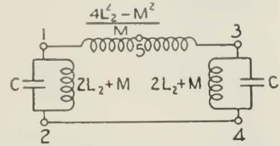


Fig. 40 Filter Section, Containing No Mutual Inductance, Equivalent to the Section of Fig. 39

This structure is listed as band pass section 1 – 4 in Table II and has propagation characteristic No. 7 of Fig. 7, and mid-shunt image impedance characteristic No. 14 of Fig. 8. Consequently, the section of Fig. 39 may be joined efficiently to any filter section of Table II having the mid-shunt image impedance characteristic No. 14 of Fig. 8 or to any section containing mutual inductance and having the same mid-shunt image impedance characteristic. The section of Fig. 39 is not capable of mid-series termination, since point 5 of inductive element 1 – 3 of Fig. 40 is not physically accessible.

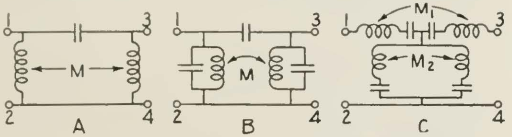


Fig. 41 Examples of Filter Sections Containing Mutual Inductance

Three further examples of the substitutions which have been discussed are represented in Figs. 41A, B, and C. By means of substitutions these structures are evidently equivalent to series-shunt sections 4 – 1 (mid-shunt terminated), 4 – 4, (mid-shunt terminated), and 3 – 7 (mid-series terminated), respectively, and they have the characteristics detailed in Table II. The above examples represent only a few of the many variant forms of structures which may be constructed by means of the various equivalences heretofore discussed.

Types of Sections Obtainable Whose Series-Shunt Equivalent Sections Contain Negative Inductances. It has already been pointed out that the following meshes of Fig. 5 may be at least partly composed of mutual inductances:—Nos. 1, 3, 4, *5a*, *5b*, *7a*, *7b*, *8a* and *8b*. When

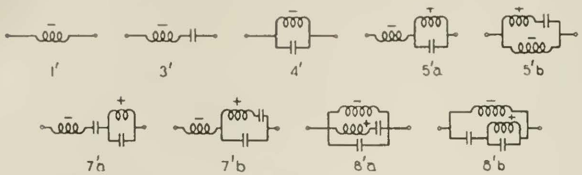


Fig. 42—Two-Terminal Reactance Meshes of Four or Less Elements, Containing Negative Inductance and Effectively Realizable Within Filter Sections

the connection of the coils is such that the mutual inductance effectively results in producing a *negative arm* in the mesh in which the mutual inductance exists, the meshes may be shown as illustrated in Fig. 42. The reactance-frequency characteristics of these arms are given in Fig. 43. It is to be noted that two general forms of reactance characteristics exist for arms *5a'* and *5b'* and that one form of reactance characteristic

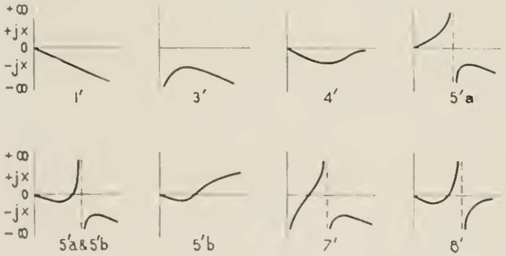


Fig. 43—Reactance-Frequency Characteristics of the Meshes of Fig. 42 Shown in Symbolic Form

is common to the two reactance arms. This duality of characteristic arises from the fact that the arms each contain two inductances, one positive and one negative, and that the general shape of the reactance characteristic is determined by the predominance of either the positive or the negative inductance. The characteristic which is peculiar to arm *5a'* occurs when the negative inductance of this arm is smaller than the positive inductance. Likewise, the characteristic peculiar to arm *5b'* occurs when the negative inductance of this arm is larger than the positive inductance. The characteristic which is common to both arms $5a'$ and $5b'$ corresponds to the alternative conditions regarding the relative magnitudes of the negative and positive inductances and the two arms $5a'$ and $5b'$ are potentially equivalent under these conditions. By means of feasible combinations of the reactance arms of Figs. 5 and 42, there can be *physically constructed* a limited number of prototype wave filter sections having no more than one transmission or one attenuation band. Such sections—involving not more than a total of six reactance elements in the series and shunt arms—are listed in Table III.

TABLE III

Tabulation of the Propagation and Impedance Characteristics of Series-Shunt Wave Filter Sections which can be Formed from the Reactance Meshes of Figs. 5 and 42

SERIES ARM

SHUNT ARM	1	3	5a	7a or 7b
1′	No Pass Band	17-13-*	21-22-*	Double Band Pass
3′	15-1-*	16-13-*	Low-and-Band Pass	Double Band Pass
5′a	16-16-*	17-17-* 18-13-* 19-13-*	16-22-* 20-22-* 21-22-* 22-22-* 32-22-*	More Than Six Elements
7′a or 7′b	Low-and-Band Pass	16-17-* 20-13-*	More Than Six Elements	More Than Six Elements

SERIES ARM

SHUNT ARM	1′	4′	5′b	8′a or 8′b
1	No Pass Band	25-*-9	24-*-16 26-*-16	High-and-Band Pass
4	23-*-14	26-*-14	23-*-19 27-*-14 28-*-14	26-*-19 31-*-14
5b	24-*-24	High-and-Band Pass	24-*-37 26-*-24 29-*-24 30-*-21 31-*-24	More Than Six Elements
8a or 8b	Double Band Pass	Double Band Pass	More Than Six Elements	More Than Six Elements

The representation of the characteristics of the structures of Table III is similar to the scheme of Table II. The figures at the top and side (for example 1 – 3′) indicate respectively, the series and shunt reactance meshes of Figs. 5 and 42 which form the prototype sections.

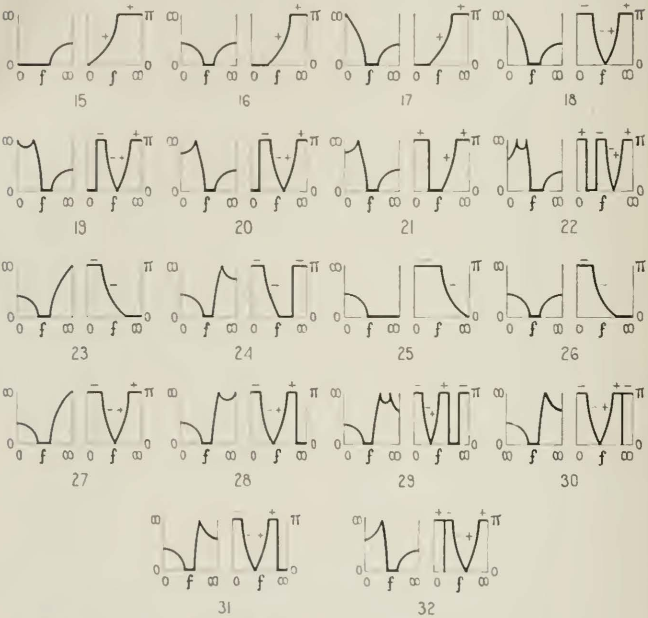


Fig. 41—Propagation Constant (Attenuation and Phase Constant) Characteristics of Filter Sections Containing Negative Inductances, Shown in Symbolic Form

The figures in the corresponding box (for example, 15 – 1 – *) indicate that the structure has propagation characteristic No. 15 of Fig. 41, and mid-series image impedance No. 1 of Fig. 8. The symbol * indicates, when inserted in the second or third position, that the structure is not physically capable of mid-series or mid-shunt termination, respectively.

It will be noted that only one low pass prototype section (1 – 3′) is given in the table, exclusive of special cases of band filter structures.

Its attenuation characteristic (No. 15 of Fig. 11) is unique as a low pass characteristic in that *the attenuation constant is finite at all frequencies.* The phase characteristic simulates, in a general way, that of the two element low pass filter (see propagation characteristic No. 1 of Fig. 7) but the phase shift in the transmission band is, in general, different. Since the structure has mid-series image impedance characteristic No. 1 it may be joined efficiently (i.e., without reflection losses) to sections of the $1-2$ and $1-3$ types.

Similarly, high pass prototype section $1'-1$ has a unique high pass attenuation characteristic in that the attenuation constant is finite at all frequencies. The phase characteristic is, in general, similar to that of the two element high pass filter $2-1$ except for the values of the phase constant in the transmission band. The section may be joined efficiently at mid-shunt to sections of the $2-1$ and $1-1$ types since it has the same mid-shunt image characteristic (No. 9).

The attenuation characteristics of the band pass prototypes listed in Table III will, in general, differ from the attenuation characteristics of structure listed in Table II. However, many of them differ only in minor respects and could have been represented identically in the symbolic fashion of Fig. 7. Inasmuch as such structures will not, however, have exactly the same attenuation characteristics for given cut-off frequencies and frequencies of infinite attenuation, different symbols or diagrams have been employed to represent them.

Certain characteristics are worthy of comment because they are not obtainable, even approximately, in structures not having negative inductance. For example, propagation characteristics Nos. 16 and 26 (Fig. 11) are band pass filter characteristics having finite attenuation at all frequencies. Characteristics No. 22 and No. 29 are unique in that there exist two frequencies of infinite attenuation, located on one side of the pass band. The attenuation constant is, in general, finite at zero and at infinite frequencies. Characteristics 19 and 28 are special cases of Nos. 22 and 29, respectively, and have two frequencies of infinite attenuation on one side of the pass band. In the case of 19, the attenuation is infinite at zero frequency and at a frequency between zero and the lower cut-off frequency. Characteristic 28 has infinite attenuation at infinite frequency and also at a frequency between the upper cut-off frequency and infinite frequency. Characteristics Nos. 18 and 27 have confluent band characteristics and have only one frequency of infinite attenuation, located either at zero frequency or at infinite frequency. Finally, characteristics Nos. 20 and 31 are confluent characteristics in each of which one frequency of infinite attenuation occurs and the attenuation is finite at zero frequency and infinite frequency.

As a general rule the phase shift characteristics shown in Fig. 44 are similar to the corresponding characteristics shown in Fig. 7. The phase characteristics of the former, within the pass bands are, in general, however, of a distinctly different character than those of the latter even though the phase constant at the cut-off frequency and the mid-frequency may be the same. Phase characteristics 21 and 24 (Fig. 44) are of special interest, however, in that while they belong to the peak type sections, the phase is of the same sign throughout the entire frequency range. Also phase characteristics 22, 29, 30 and 32 have a unique property, for band pass structures, in that the phase undergoes a change in sign within one attenuation band.

In regard to the impedance characteristics, it is noted from Table III that *no novel impedance characteristics are obtained in structures having negative inductances as compared to the structures not having negative inductances.* This is a valuable property of the prototype structures listed in Table III as it permits composite filters to be readily formed utilizing both the sections of Tables II and III.[16]

Characteristics of a Typical Filter. In order to illustrate the derivation of design formulae for a specific prototype having negative inductances, consider as an example the band pass structure $3-3'$ of Table III. We shall neglect the effect of dissipation on the characteristics of the structure, as the treatment of dissipation has been previously outlined. The prototype cited is illustrated in Fig. 45A. Two

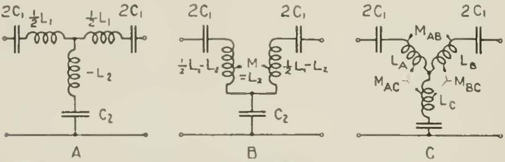


Fig. 45—Prototype Section Containing Negative Inductance, and Two of Its Physically Realizable Forms

methods of physically obtaining such a prototype are illustrated in Figs. 45B and C. In this structure the series impedance Z_1 is

$$Z_1 = j\left(\omega L_1 - \frac{1}{\omega C_1}\right). \tag{92}$$

[16] For a general method of proving the equality of the image impedances of sections containing negative inductance and of appropriate sections containing no negative inductance, refer to the Appendix.

The impedance of the shunt arm is

$$Z_2 = -j\left(\omega L_2 + \frac{1}{\omega C_2}\right). \tag{93}$$

The ratio, $Z_1/4Z_2$, which controls the attenuation and phase constants, per section, of the structure is

$$\frac{Z_1}{4Z_2} = \frac{j\left(\omega L_1 - \dfrac{1}{\omega C_1}\right)}{-j\left(\omega L_2 + \dfrac{4}{\omega C_2}\right)} = \frac{C_2}{4C_1}\,\frac{1 - L_1C_1\omega^2}{1 + L_2C_2\omega^2}. \tag{94}$$

From the impedance characteristics of reactance meshes 3 and 3′, as illustrated in Figs. 6 and 43, and the combined reactance character-

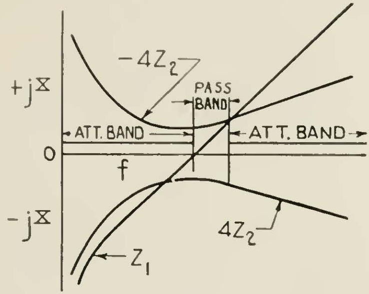


Fig. 46—Reactance-Frequency Characteristics of the Series and Shunt Arms of the Prototype Section of Fig. 45-A

istics of Fig. 46 for Z_1, $4Z_2$ and $-4Z_2$, it will be noted that the lower cut-off frequency, f_1, is that at which $Z_1 = 0$. Hence,

$$f_1 = \frac{1}{2\pi\sqrt{L_1C_1}}. \tag{95}$$

Similarly, the upper cut-off frequency is that at which $Z_1 = -4Z_2$ or $j\omega L_1 - j/\omega C_1 = j4\omega L_2 + j4/\omega C_2$. From this relationship, the upper cut-off frequency is

$$f_2 = \frac{1}{2\pi}\sqrt{\frac{C_2 + 4C_1}{C_1C_2(L_1 - 4L_2)}}. \tag{96}$$

Let f_r be assumed as the frequency where Z_2 is a minimum, that is, where $\omega^2 L_2C_2 = 1$. We may then write

$$f_r = \frac{1}{2\pi\sqrt{L_2C_2}}. \tag{97}$$

Substituting the above values of f_1, f_2 and f_r in formula (91) we obtain for $Z_1/4Z_2$

$$\frac{Z_1}{4Z_2}=\frac{\left[1-\left(\frac{f}{f_1}\right)^2\right]\left[\left(\frac{f_2}{f_r}\right)^2+1\right]}{\left[1+\left(\frac{f}{f_r}\right)^2\right]\left[\left(\frac{f_2}{f_1}\right)^2-1\right]}. \tag{98}$$

From this last expression the attenuation and phase characteristics may be plotted from formulae (22) and (23) or from Figs. 11 and 12. The attenuation and phase constant characteristics are shown symbolically as characteristic 16 of Fig. 41. This structure has unusual attenuation properties which have already been discussed.

From equation (6) and the values of Z_1 and Z_2, in (92) and (93), the mid-series image impedance (Z_o), at the mid-frequency, is

$$Z_o=\frac{1}{2}\left[\sqrt{\frac{L_1}{C_1}+\frac{4L_1}{C_2}}-\sqrt{\frac{L_1}{C_2}-\frac{4L_2}{C_1}}\right]. \tag{99}$$

Since the mid-series image impedance, at any frequency, is the same as that of filter section 3 − 3, we have:

$$Z_I=Z_o\sqrt{1-\frac{\left[\frac{f}{f_m}-\frac{f_m}{f}\right]^2}{\left[\frac{f_2}{f_m}-\frac{f_m}{f_2}\right]^2}}=Z_o\sqrt{1-\frac{\left[\frac{f}{\sqrt{f_1f_2}}-\frac{\sqrt{f_1f_2}}{f}\right]^2}{\left[\sqrt{\frac{f_2}{f_1}}-\sqrt{\frac{f_1}{f_2}}\right]^2}} \tag{100}$$

where f_m is the mid-frequency ($f_m=\sqrt{f_1f_2}$), as before.

The prototype is not capable of mid-shunt termination, hence, its hypothetical mid-shunt impedance characteristic will not be derived.

From the preceding formulae, explicit expressions may be derived for the values of L_1, C_1, L_2 and C_2

$$L_1=\frac{Z_o m'}{\pi(f_2-f_1)}, \tag{101}$$

$$C_1=\frac{f_2-f_1}{4\pi f_1^2 Z_o m'}, \tag{102}$$

$$L_2=\frac{-Z_o}{\pi(f_2-f_1)}\cdot\frac{1-m'^2}{4m'}, \tag{103}$$

$$C_2=\frac{(f_2-f_1)m'}{\pi Z_o(f_2^2-f_1^2m'^2)}, \tag{104}$$

$$m'=\sqrt{1+\frac{\left(\frac{f_2}{f_1}\right)^2-1}{\left(\frac{f_r}{f_1}\right)^2+1}}. \tag{105}$$

As a numerical example of the solution of the prototype discussed assume, as in the example following equation (11), that the lower cut-off frequency f_1 is 20,000 cycles and that the upper cut-off frequency f_2 is 25,000 cycles. Assume f_r, a convenient parameter for the families of attenuation and phase constant curves which this section may have, for any given cut-off frequency, to be 30,000 cycles. Assume that the value of the mid-series image impedance Z_0 at the mid-frequency is 600 ohms; then from formula (99) $m' = 1.083$: hence $L_1 = .0412$ henries, $C_1 = .00153 \times 10^{-6}$ farads, $L_2 = .00152$ henries and $C_2 = .0184 \times 10^{-6}$ farads. The structure with the numerical values of inductance and capacity for this specific example is shown in Fig. 17A.

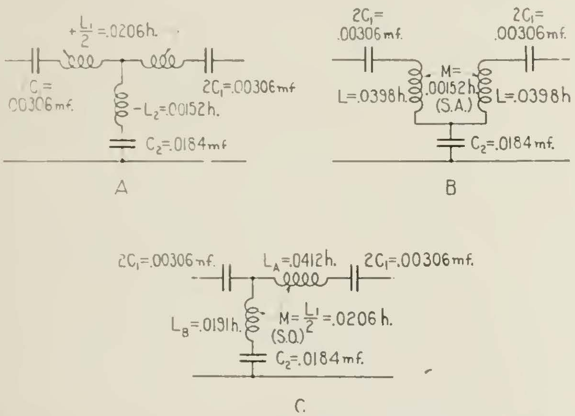
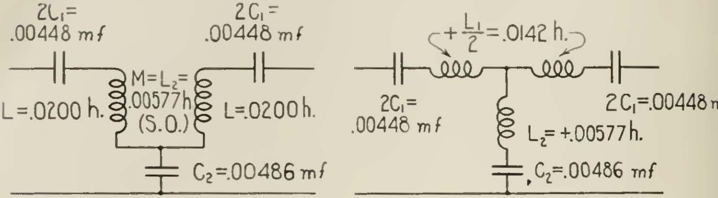


Fig. 47—Numerical Example of Equivalent Filter Sections Containing Negative Inductance

If, for the T mesh inductances in Fig. 47A, we substitute a transformer mesh having the values shown in Fig. 17 B, the mesh of the latter figure is electrically equivalent to the prototype structure and is an example of the method of employing the structure. Similarly, Fig. 47C illustrates the substitution of another type of three element mesh for the coil mesh of the prototype structure of Fig. 17A and is another example of the manner in which the prototype may be physically expressed.

The structure of Fig. 17B represents a similar case to that of 48A. However, as the mutual inductance is here series opposing, the prototype series-shunt equivalent structure is shown in Fig. 48B and contains no negative inductances. It will be found that the values chosen correspond to the numerical example of the structure 3—3 following equation 41.



A B

Fig. 48—Numerical Example of a Filter Section Containing No Negative Inductance

APPENDIX

Conditions for the Equality of the Image Impedances of Typical Filter Structures

It has been stated that the formation of recurrent and composite wave filters is dependent upon the maintenance of equal image impedance characteristics (of the sections or half-sections joined) at each junction point throughout the filter.

A general method of ascertaining the conditions for the equality of image impedance characteristics will be demonstrated by illustrations from typical pairs of sections.

Illustration No. 1—Negative Inductance in Shunt Arm of One Structure. Consider the filter sections listed as 3–4 (confluent structure) in Table II, and 3–1′ in Table III. It will be shown that, under proper conditions, their mid-series image impedance characteristics may be made equal at all frequencies. (By reference to the above tables, both sections have mid-series impedance characteristic No. 13 of Fig. 8).

From equation (6)

$$Z_I^2 = Z_1Z_2 + \frac{Z_1^2}{4}. \tag{106}$$

In Fig. 19, let

$$Z_1 = Z_{1A} + Z_{1B} = j\omega L_1 + \frac{1}{j\omega C_1}, \tag{107}$$

$$Z_1' = K_A Z_{1A} + K_B Z_{1B}, \tag{108}$$

and $$Z_2' = -K_C Z_{1A}, \tag{109}$$

where $K_A = L_1'/L_1$, $K_B = C_1/C_1'$ and $K_C = L_2'/L_1$. (110)

From (106)

$$Z_I^2 = R^2 + \frac{Z_1^2}{4} \tag{111}$$

in which

$$R = \sqrt{\frac{L_2}{C_1}} = \sqrt{\frac{L_1}{C_2}}. \tag{112}$$

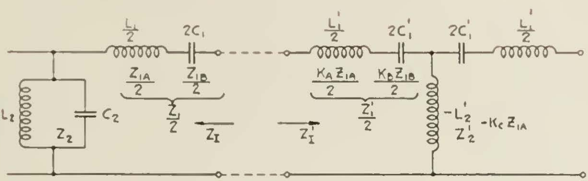


Fig. 49—Two Structures Having Equal Mid-Series Image Impedances, One of Which Contains a Negative Inductance in Its Shunt Arm

From (107) and (111)

$$Z_I^2 = R^2 + \tfrac{1}{4}(Z_{1A} + Z_{1B})^2 = 1/4\, Z_{1A}^2 + (1 + K/2)R^2 + 1/4\, Z_{1B}^2 \tag{113}$$

where $K = Z_{1A}Z_{1B}/R^2 = L_1/L_2 = C_2/C_1$. (114)

Now from (106) and (108)

$$(Z_I')^2 = Z_1'Z_2' + \frac{(Z'_1)^2}{4} = \left(\frac{K_A^2}{4} - K_A K_C\right) Z_{1A}^2 + \left(\frac{K_A K_B}{2} - K_B K_C\right) K R^2 + \frac{K_B^2}{4} Z_{1B}^2. \tag{115}$$

Since, by postulation, in Fig. 49, $Z_I = Z_I'$, we may equate the coefficients of (113) and (115). This gives

$$\frac{1}{4} = \frac{K_A^2}{4} - K_A K_C, \tag{116}$$

$$1 + \frac{K}{2} = \left(\frac{K_A K_B}{2} - K_B K_C\right) K, \tag{117}$$

and $$\frac{1}{4} = \frac{K_B^2}{4}. \tag{118}$$

Whence
$$K_B = \frac{C_1}{C_1'} = 1, \tag{119}$$

and
$$K_A = \frac{L_1'}{L_1} = \frac{L_1'C_1'}{L_1C_1} = \frac{f_M^2}{f_1^2} = \frac{f_2}{f_1}, \tag{120}$$

where f_1 and f_2 are the lower and upper cut-off frequencies, respectively, and $f_M = \sqrt{f_1 f_2}$ of the structures of Fig. 49.

From (116) and (120)

$$K_C = \frac{L_2'}{L_1} = \frac{1}{4}\left(K_A - \frac{1}{K_A}\right) = \frac{1}{4}\left(\frac{f_2}{f_1} - \frac{f_1}{f_2}\right). \tag{121}$$

Therefore, when the relationships between the constants of the two structures of Fig. 49 satisfy equations (119), (120) and (121), the structures will have the same mid-series image impedance characteristics. Explicit relations for the values of C_1', L_1' and L_2' may be obtained from equations (119), (120) and (121) as follows:

$$C_1' = C_1, \tag{122}$$

$$L_1' = L_1\frac{f_2}{f_1}, \tag{123}$$

$$L_2' = \frac{L_1}{4}\left(\frac{f_2}{f_1} - \frac{f_1}{f_2}\right). \tag{124}$$

Consequently, if the constants and cut-off frequencies of a confluent structure are known, the constants of a structure of the 3-1′ form having an identical mid-series image impedance characteristic can be derived from equations (122), (123) and (124).

Illustration No. 2 Negative Inductance in Series Arm of One Structure. Consider next the filter sections listed as 3-4 (confluent structure) in Table II and 1′-4 in Table III. It will be shown that, under proper conditions, their mid-shunt image impedance characteristics may be made equal at all frequencies. (By reference to the above tables, both sections have mid-shunt impedance characteristic No. 11 of Fig. 8).

From equation (7

$$Y_I^2 = Y_1 Y_2 + \frac{Y_2^2}{4}, \tag{125}$$

where $Y_1 = 1/Z_1$, $Y_2 = 1/Z_2$ and $Y_I = 1/Z_I$.

In Fig. 50, let

$$Y_2 = Y_{2A} + Y_{2B} = \frac{1}{j\omega L_2} + j\omega C_2. \tag{126}$$

$$Y_2' = K_A Y_{2A} + K_B Y_{2B}, \tag{127}$$

and

$$Y_1' = -K_C Y_{2A}, \tag{128}$$

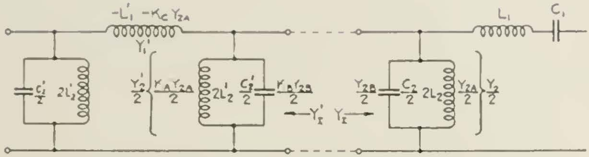


Fig. 50—Two Structures Having Equal Mid-Shunt Image Impedances, One of Which Contains a Negative Inductance in Its Series Arm

where $K_A = L_2/L_2'$, $K_B = C_2'/C_2$ and $K_C = L_2/L_1'$. (129)

From (125)

$$Y_I^2 = G^2 + \frac{Y_2^2}{4} \tag{130}$$

in which

$$G = \sqrt{\frac{C_1}{L_2}} = \sqrt{\frac{C_2}{L_1}} \tag{131}$$

From (126) and (130)

$$Y_I^2 = G^2 + 1/4 (Y_{2A} + Y_{2B})^2 = 1/4\, Y_{2A}^2 + (1 + K/2) G^2 + 1/4\, Y_{2B}^2 \tag{132}$$

where

$$K = Y_{2A} Y_{2B}/G^2 = L_1/L_2 = C_2/C_1. \tag{133}$$

Now from (125) and (127)

$$(Y_I')^2 = Y_1' Y_2' + \frac{(Y_2')^2}{4} = \left(\frac{K_A^2}{4} - K_A K_C\right) Y_{2A}^2 +$$

$$\left(\frac{K_A K_B}{2} - K_B K_C\right) K G^2 + \frac{K_B^2}{4} Y_{2B}^2 \tag{134}$$

Since, by postulation, in Fig. 50, $Y_I = Y_I'$, we may equate the coefficients of (132) and (134). This gives

$$\frac{1}{4} = \frac{K_A^2}{4} - K_A K_C, \tag{135}$$

$$1 + \frac{K}{2} = \left(\frac{K_A K_B}{2} - K_B K_C\right) K, \tag{136}$$

and $$\frac{1}{4} = \frac{K_B^2}{4}. \tag{137}$$

Whence $$K_B = \frac{C_2'}{C_2} = 1. \tag{138}$$

and $$K_A = \frac{L_2}{L_2'} = \frac{L_2 C_2}{L_2' C_2'} = \frac{f_2^2}{f_M^2} = \frac{f_2}{f_1} \tag{139}$$

where f_1 and f_2 are the lower and upper cut-off frequencies, respectively, and f_M is the mean frequency ($\sqrt{f_1 f_2}$) of the structures of Fig. 50.

From (135) and (139)

$$K_C = \frac{L_2}{L_1'} = \frac{1}{4}\left(K_A - \frac{1}{K_A}\right) = \frac{1}{4}\left(\frac{f_2}{f_1} - \frac{f_1}{f_2}\right). \tag{140}$$

Therefore, when the relationships between the constants of the two structures of Fig. 50 satisfy equations (138), (139) and (140), the structures will have the same mid-shunt image impedance characteristics. Explicit relations for the values of C_2', L_2' and L_1' may be obtained from equations (138), (139) and (140) as follows:

$$C_2' = C_2, \tag{141}$$

$$L_2' = L_2 \frac{f_1}{f_2}, \tag{142}$$

$$L_1' = \frac{4L_2}{\left(\frac{f_2}{f_1} - \frac{f_1}{f_2}\right)}. \tag{143}$$

Therefore, if the constants and cut-off frequencies of a confluent structure are known, the constants of a structure of the $1'-4$ form having an identical mid-shunt image impedance characteristic can be derived from equations (141), (142) and (143).

BIBLIOGRAPHY

1. Thévenin, M. L., "Sur un Nouveau Théorème d'Electricité Dynamique," *Comptes Rendus*, Vol. 97, pp. 159–161 (1883).
2. Kennelly, A. E., "The Equivalence of Triangles and Three-Pointed Stars in Conducting Networks," *Electrical World and Engineer*, New York, Vol. XXXIV, pp. 413–414, Sept. 16, 1899.
3. Campbell, G. A., "Cisoidal Oscillations," *Trans. A. I. E. E.*, Vol. XXX, Part II, pp. 873–909 (1911).
4. Campbell, G. A., U. S. Patents Nos. 1,227,113 and 1,227,114 (1917).
5. Gherardi, B. and Jewett, F. B., "Telephone Repeaters," *Trans. A. I. E. E.*, 1919.
6. Wagner, K. W., *Arch. fur Elektrotechnik*, Vol. 8, p. 61 (1919); *E. T. Z.*, Aug. 7, 1919.

7. Van der Bijl, H. T., "Thermionic Vacuum Tubes," Published 1920.
8. Pierce, G. W., "Electric Oscillations and Electric Waves," Published, 1920.
9. Colpitts, E. H. and Blackwell, O. B., "Carrier Current Telephony and Telegraphy," *Trans. A. I. E. E.*, Feb., 1921.
10. Clement, L. M., Ryan, F. M., and Martin, D. K., "The Avalon-Los Angeles Radio Toll Circuit," *Proc. I. R. E.*, May, 1921.
11. Fletcher, H., "The Nature of Speech and Its Interpretation," *Jour. Franklin Inst.*, June, 1922.
12. Campbell, G. A., "Physical Theory of the Electric Wave-Filter," *Bell Sys. Tech. Jour.*, Nov., 1922.
13. Zobel, O. J., "Theory and Design of Uniform and Composite Electric Wave-Filters," *Bell Sys. Tech. Jour.*, Jan., 1923.
14. Rose, A. F., "Practical Application of Carrier Telephone and Telegraph in the Bell System," *Bell Sys. Tech. Jour.*, April, 1923.
15. Hartley, R. V. L., "Relation of Carrier and Side-Bands in Radio Transmission," *Bell Sys. Tech. Jour.*, April, 1923.
16. Bown, C. D., Englund, C. R., and Friis, H. T., "Radio Transmission Measurements," *Proc. I. R. E.*, April, 1923.
17. Peters, L. J., "Theory of Electric Wave Filters Built up of Coupled Circuit Elements," *Jour. A. I. E. E.*, May, 1923.
18. Demarest, C. S., "Telephone Equipment for Long Cable Circuits," *Bell Sys. Tech. Jour.*, July, 1923.
19. Nichols, H. W. and Espenschied, L., "Radio Extension of the Telephone System to Ships at Sea," *Bell Sys. Tech. Jour.*, July, 1923.
20. Carson, J. R. and Zobel, O. J., "Transient Oscillations in Electric Wave-Filters," *Bell Sys. Tech. Jour.*, July, 1923.
21. Arnold, H. D. and Espenschied, L., "Transatlantic Radio Telephony," *Bell Sys. Tech. Jour.*, Oct., 1923.
22. Best, F. H., "Measuring Methods for Maintaining the Transmission Efficiency of Telephone Circuits," *Jour. A. I. E. E.*, Feb., 1924.
23. Casper, W. L., "Telephone Transformers," *Jour. A. I. E. E.*, March, 1924.
24. Slaughter, N. H. and Wolfe, W. V., "Carrier Telephony on Power Lines," *Jour. A. I. E. E.*, April, 1924.
25. Foster, R. M., "A Reactance Theorem," *Bell Sys. Tech. Jour.*, April, 1924.
26. Martin, W. H., "The Transmission Unit and Telephone Transmission Reference System," *Bell Sys. Tech. Jour.*, July, 1924.
27. Zobel, O. J., "Transmission Characteristics of Electric Wave-Filters," *Bell Sys. Tech. Jour.*, Oct., 1924.

Some Contemporary Advances in Physics—VI
Electricity in Gases

By KARL K. DARROW

1. Introduction

THE physicists of a quarter of a century ago, who devoted themselves to the study of electricity in gases, were happily inspired; for among the myriad of intricate and obscure phenomena which they observed there are some few of an extreme simplicity, in which the qualities of the individual atoms of matter and electricity are manifest; in analyzing these they entered upon the path that led most directly to the deeper understanding of nature which is superseding the physics of the nineteenth century, and the physics of today is founded upon their efforts. The electron was perceived for the first time in the course of observations on the electric discharge in rarefied gases, and other experiments in the same field established the atom in science as a real and definite object. The discovery of the atom is commonly credited to the chemists; yet fifteen years have not passed since students of chemistry were being warned by a famous teacher that "atom" and "molecule" are figurative words, not on any account to be taken literally! The laws of chemical combination were held insufficient to prove that atoms have any real existence; though elements may always combine with one another in unchanging proportions, this does not prove anything about the weights of the atoms, or their sizes, or their qualities, or even that all the atoms of an element have the same weight, or even that there are any atoms at all. Now that we are past the necessity for this caution, and can count atoms, and measure their masses, and infer something about their structure, and estimate how close together they can approach, and know what happens to them when they strike one another or are struck by electrons; now that we can fill in the picture of the atom with so many and so diverse details, we are indebted for this progress chiefly to the men who gathered the data and made the theories concerning the conduction of electricity in gases. Many will remember how in the years before the great war this field of research seemed the most vital part of physics, the most inspired with a sense of new life and swift advance; now others share with it the centre of the stage, but they won their places chiefly because of the light it shed upon them.

It seems strange that the flow of electricity in gases should have proved easier to interpret than the flow of electricity in metals, which in appearance is certainly by far the simpler. One applies the terminals of a battery to the ends of a wire, and promptly the electric potential distributes itself with a uniform gradient along the wire and a current flows steadily down it. So rigorously is the current proportional to the voltage between the ends of the wire, over very wide ranges of voltage and current, that we regard the ratio as an essential constant of the wire; and we regard the ratio of potential-gradient (electric field) to current density as an essential characteristic of the metal, and give it a name resistivity or specific resistance and refer to theories of conduction in metals as theories of metallic resistance. It all seems exceedingly simple, and yet in the foregoing article of this series I have shown how all the attempts to interpret it have gone in vain. Much more complex in appearance is the discharge through a gas. One applies the terminals of a battery to a pair of electrodes facing one another in the open air, and perhaps nothing happens, or so minute a current flows that the most delicate of instruments is demanded to detect it; and then when the battery-voltage is very slightly raised, there may be an explosion with a blaze of light, dissociating the gas and corroding the electrodes, and draining off the available electricity in a moment. Or if one of the electrodes is acutely pointed there may be glows and luminous sheaths around it or tentacles of bluish light ramifying from it far and wide through the air. Or the discharge may rise to the heat of incandescence, and the gas and the electrodes shine with a blinding radiance, the brightest light that can be kindled on the earth. Or if the electrodes are enclosed in a tube containing a rarefied gas or vapor, the gas flares up into an extraordinary pattern of light and shade, lucent vividly-colored clouds floating between regions glowing feebly or obscure; and as the gas is gradually pumped away, the pattern changes and fades, a straight beam of electrons manifests itself by a luminous column traversing the tube, the glass walls flash out in a green fluorescence, and finally all becomes extinct. As for that even gradient, and that constant proportion between current and field strength distinguishing the metals, we cannot find them here. There is no such thing as the resistance of a gas; we had better forget the word, we cannot attach any physical meaning to the ratio of current and voltage.

I must not give the impression that all these manifold forms of the electric discharge in gases are understood. Certain of the simplest of them have been clarified, and as a result still simpler ones have been realized and comprehended in their turns, and so on down to the simplest of all, which is the discharge across a vacuum. This sounds somewhat like a paradox and so it would have seemed thirty or forty years ago, when electricity was thought to be inseparable from matter, and the only known discharges across gases were the discharges in which the gas plays an indispensable role. It is important to note the manner of this evolution, for much of the history of modern physics is dominated by it. We should not be nearly so far advanced as we are, had we not learned two things: how to reduce the amount of gas in a tube until an electron can fly clear across it with scarcely any chance of meeting an atom, and how to persuade an electron to emerge from a metal otherwise than by starting a discharge in a gas over its surface. We who are so familiar with the idea of electrons boiling out of a hot wire, or driven out of a cold metal plate by light shining upon it, or fired as projectiles out of exploding atoms, find it difficult to imagine the confusion which of necessity prevailed when all these processes were unknown. In the early stages of research into the discharge in gases, it was made clear that of each self-maintaining discharge a stream of electrons flowing out of the negative electrode is an essential part; the electron-stream maintains the gas-discharge, and reciprocally the gas-discharge maintains the electron-stream. The latest stage commenced when it was made possible to produce and maintain such an electron-stream independently of any gas-discharge, and deal with it at will.

Let me then begin the exposition with this idea, which so many years of research were required to render acceptable: the idea of a stream of electrons emerging from a metal wire or a metal plate, at a constant rate which is not influenced by the presence or absence of gas in the space surrounding the metal. The reader may think either of thermionic electrons flowing spontaneously out of a hot wire, or of photo-electrons flying out of a metal plate upon which ultra-violet light is shining.[1]

2. The Flow of Electrons Through a Very Rarefied Monatomic Gas, and Their Encounters with the Atoms

Conceive a source of electrons, a negative electrode or cathode, which is enclosed in a tube. If the tube is highly evacuated, the

[1] While forming one's ideas it is preferable to think of the photoelectric source, for a variety of reasons; the electron-stream is not very dense, the electrons emerge with kinetic energies never in excess of a certain sharply-marked limiting value, the metal is cold and not likely to react chemically with whatever gas surrounds it. Also several of the classical fundamental experiments were performed in the years from 1898 to 1906, when the photoelectric effect had become a reliable instrument of research and the thermionic effect had not. Nowadays it is sometimes used in the hope of surpassing the accuracy of earlier work, or in experiments on compound gases which the hot wire might decompose. Still the hot wire is so much easier to insert and handle, its emission so much more convenient and controllable, that it will no doubt be employed in the great majority of experiments in the future as in the past.

electrons enter the vacuum freely; electricity has no horror of a vacuum, any more than nature generally. Still there is something which suggests the *horror vacui* of the scientists before Galileo; for the electrons which are already partway across the vacuum tend, by their electrostatic repulsion, to push back their followers which are just emerging from the metal. This is the space-charge effect, which has become famous since the audion became almost as common an object as the incandescent lamp in the American home. I shall presently have to write down the equations describing this effect; for the time being we may ignore it, so long as the electron-stream is not more profuse than a photoelectric current generally is. The electrons of these scanty discharges enter into the vacuum and pass over without hindrance.

At this point it is advisable to say what is meant by a "vacuum." Scientists are growing more exigent year by year in their use of this term; thirty or forty years ago people spoke of "vacuum tubes" meaning tubes so full of gas that they would transmit a big current with a resplendent luminous display, but this self-contradicting usage has become quite intolerable. At the present day the least density of gas, or the highest vacuum, commonly attained corresponds to a gas-pressure about 10^{-11} as great as the pressure and density of the atmosphere. This means that there are about 10^{8} molecules in a cubic centimetre of the "vacuum," which may make the name sound absurd. But the practical criterion for a vacuum is not whether the remaining atoms seem many or few, but whether they are numerous enough to affect the passage of a discharge; and as an electron shooting across a tube 10 cm. wide and evacuated to this degree has 999999 chances out of a million of getting clear across without encountering a molecule, the tube is vacuous enough for any sensible definition.

Next we will imagine that a gas is introduced into the tube, in quantity sufficient so that each electron going from cathode toward anode will collide on the average with one or possibly two atoms on its way. It is best to begin by thinking of one of the noble gases, of which helium, argon and neon are the ones in common use; or of the vapour of a metal, mercury vapour being much the easiest of these to work with; for their atoms behave in a simpler and clearer manner toward the electrons than do the molecules of the commonest gases, particularly the oxygen molecules which are so numerous in air. In fact the practice of using the noble gases and the metal vapours—that is to say, the *monatomic* gases—wherever possible in these researches ought really to be regarded as one of the great advances of the last few years; our predecessors would certainly have learned more about the discharge in gases than they ever did, if they had not studied it in air ninety times out of a hundred, and in other diatomic gases most of the other ten.

Let us suppose that the tube contains helium of the extremely small density I have just defined. Then so long as the kinetic energy of an electron does not exceed 19.75 volts, it will rebound from any helium atom which it strikes, like a very small perfectly elastic ball rebounding from a very large one. We might conceive the contents of the tube (for this purpose and only for this purpose!) as a flock of immense ivory pushballs floating languidly about, with a blizzard of equally elastic golfballs or marbles darting through the interspaces and occasionally striking and bouncing off from one of the pushballs. If the collisions between electrons and atoms are perfectly elastic, as I have said without giving evidence, the electron will lose an extremely small part of its kinetic energy at each collision, owing to the great disparity in masses—a fraction varying from zero up to not more than .000537 depending on the direction of rebound.

This was verified in a pretty experiment by K. T. Compton and J. M. Benade, who utilized a certain effect[2] which electrons produce when they have kinetic energy exceeding 19.75 volts at the moment of a collision with a helium atom. For example, when the pressure of helium was 4.34 mm. and the electrons were drawn from a cathode to an anode 0.265 cm. away, a voltage-difference of 20.25 (plus an unknown correction) was required to produce this effect; when the anode was 0.90 cm. from the cathode the required voltage-difference was 23.45 (plus the same correction). The extra volts were spent in replacing the energy lost by the electrons in the collisions with helium atoms over the extra 6.3 mm.; they amounted to an average of .0003 of the electron's energy lost in each collision, excellently in agreement with the assumption.

Now as for the transit of the electron-stream from cathode to anode, the helium atoms will simply thin it down by intercepting some of the electrons and turning their courses backwards or aside. The greater the number of atoms in the path, the greater the proportion of electrons intercepted; it can easily be seen that, so long as the gas is not denser than I have specified, this proportion increases as an exponential function of the number of atoms between cathode and anode,[3] whether this number be increased by introducing more gas or by moving the anode farther away from the cathode. If

[2] Incipient ionization, as described below.

[3] The proportion increases more slowly when there are already so many atoms between anode and cathode that an electron is likely to strike two or more on its way across.

the anode and the cathode are two parallel plates d centimetres apart, and there are P helium atoms in a cubic centimetre of the gas between, and N_0 electrons start out in a second directly towards the anode from any area of the cathode, the proportion $\Delta N/N_0$ of electrons which are intercepted before they reach the anode is

$$\Delta N/N_0 = 1 - e^{-APd} \tag{1}$$

and the number of electrons reaching the corresponding area on the anode in a second, $N_0 - \Delta N$, conforms to the equation:

$$\log_e (N_0 - \Delta N) = -APd + \text{const.} \tag{2}$$

The coefficient A is a constant to be interpreted as the effective cross-sectional area of the helium atom relatively to an oncoming electron—that is, the atom behaves towards the electron like an obstacle presenting the impenetrable area A to it.

In the experiments performed to verify these assertions and determine the value of A, the simple geometrical arrangement which I have described is generally modified in one way or another for greater accuracy or convenience. Mayer approached most nearly to the simple arrangement; in his apparatus (Fig. 1) the electrons

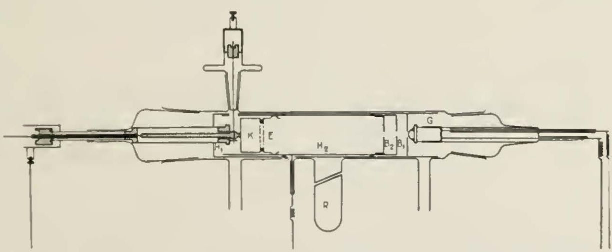


Fig. 1—Apparatus for determining the percentage of electrons which go across a gas of variable thickness without interception. (Mayer, *Annalen der Physik*)

which emerge from the hot filament at G, pass through the two slits in front of it, and then go down the long tube to the anode K, which is drawn backward step by step. The logarithmic curves of current versus distance for various pressures of nitrogen (Fig. 2) are straight. Unfortunately the current also diminishes as the distance is increased when the nitrogen is pumped out altogether; this is attributed partly to residual vapors and partly to the electrons striking the walls of the tube. The other curves are corrected for this effect, and then A is calculated. For helium it is 25.10^{-16} cm²; the values obtained by modifications of the method agree well.[4]

The helium atoms therefore behave as so many minute and yet appreciable obstacles to the passage of the electron-stream, so long as the electrons are not moving so rapidly that their energies of motion do not surpass 19.75 volts. Electrons as slow as these bounce off from the atoms which they strike. When, however, an electron possessing kinetic energy greater than 19.75 volts strikes a helium atom,

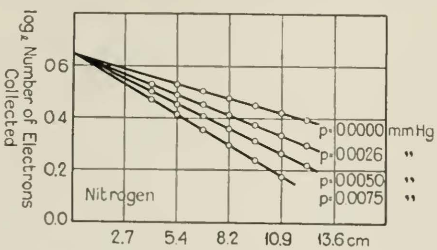


Fig. 2—Curves illustrating the interception of electrons by nitrogen molecules which they strike. (Mayer, *Annalen der Physik*)

it is liable to lose 19.75 volts of its energy to the atom, retaining only the remainder. This energy does not become kinetic energy of the atom, a process which would be incompatible with conservation of momentum; neither is the atom broken up; it receives the quota of energy into its internal economy, where some kind of a domestic change occurs with which we are not concerned for the moment, except in that it furnishes an exceedingly accurate indirect way of calculating the exact amount of energy taken from the electron. The atom is said to be put into an "excited" or sometimes into a "meta-

[4] The modified methods are generally more accurate. Ramsauer's device, which I described in the first article of this series, is probably the best. By a magnetic field he swung a stream of electrons around through a narrow curving channel, and those which were deviated even through a few degrees struck the limiting partitions and were lost from the beam; he varied the number of atoms in the channel by varying the gas-pressure. In this way he discovered that A for argon atoms differs very greatly for different speeds of the electrons; it was later found that other kinds of atoms have a variable A, although happily the variations are not great. This seems strange at first, but it is probably stranger that A should have nearly the same value for different speeds of the oncoming electrons, as for many atoms it does; and stranger yet that it should have the same value for an oncoming atom as for an oncoming electron, as is often tacitly assumed, and not too incorrectly.

stable" state, and the energy which it takes up, measured in volts, is called its *resonance-potential*. The electron is left with only the difference between its initial energy and the 19.75 volts which it surrendered.

This loss of energy in a so-called "inelastic" collision can be dem-

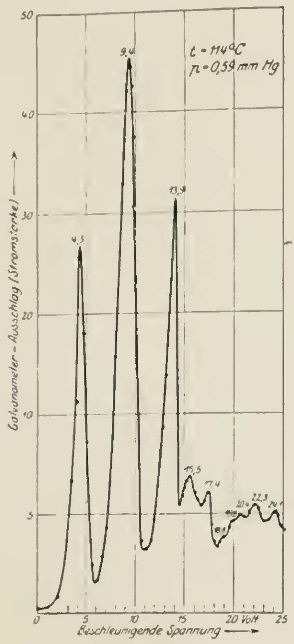


Fig. 3—Curve displaying resonance-potentials of mercury.
(Einsporn, *Z.S. f. Physik*)

onstrated by inserting a third electrode into the path of the electrons, charged negatively to just such a degree that an electron retaining its full initial speed can overcome the repulsion of the electrode and win through to it, while an electron which has lost a quantity of its kinetic energy in an inelastic collision cannot quite "make the grade." When the energy of the electrons streaming into the helium is raised just past 19.75 volts there is a sudden falling-off in the number of electrons arriving at the third electrode. The curve in Fig. 3, obtained by Einsporn, shows the current into such an electrode in mercury-vapor rising and falling again and again as the voltage passes through the values which are integer multiples of 4.9 volts, the least resonance-potential of mercury. Helium has a second resonance-potential, at 20.45 volts; neon has two, at 16.65 and 18.45 volts respectively; argon three, at 11.55, 13.0 and 14.0 volts;[5] mercury two, at 4.9 and 6.7 volts. It is almost certain that in each case these are only the most conspicuous among many, but the lowest mentioned is the lowest of all.

Up to this point we find the gas acting as a mere inert obstruction to the discharge; every collision of an electron with an atom interrupts the progress of the electron toward the anode and to that extent impedes the discharge. Past the resonance-potential the same action continues, although the interruption is doubtless less severe when the electron is deprived of part of its energy of forward motion than when it is flung backward with its motion reversed in direction and its energy intact. At the resonance-potential, the gas does begin to assist the discharge in an indirect way. Atoms which are put into an "excited state" by a blow from an electron revert of themselves to the normal state, some time later; in so doing they emit radiation, some of which falls upon the cathode; some of this is absorbed in the cathode metal, and expels electrons which go along with the maintained electron-stream as extra members of it. Thus the gas helps in increasing and maintaining the discharge; this effect is of great theoretical importance, and I will return to it later; but in these actual circumstances it is not very prominent.

The really powerful cooperation of the gas in the discharge commences when the electrons are given so great an energy that they disrupt the atoms which they strike, tearing off an electron from each and leaving a positively-charged residue, an *ion* which wanders back towards the cathode while the newly-freed electron and its liberator go on ahead towards the anode. The onset of this ionization may be detected by inserting a third electrode into the gas, it being charged negatively to such a degree that no electrons can reach it, but only positive ions; or by the increase in the current between cathode and anode, for the current increases very suddenly and very rapidly when the energy of the primary electrons is raised past the threshold-value, the *ionizing-potential* of the gas; 21.5 volts for helium, 21.5 for neon, 15.3 for argon, 10.4 for mercury. Consider for example the

[5] I take the values for neon and argon from Hertz' latest publication.

precipitate upward rush of the current-voltage curve in Fig. 4, from the work of Davis and Goucher.[6]

At this point I will digress to speak very briefly of the succession of events which occurs when the electron-stream is much denser than

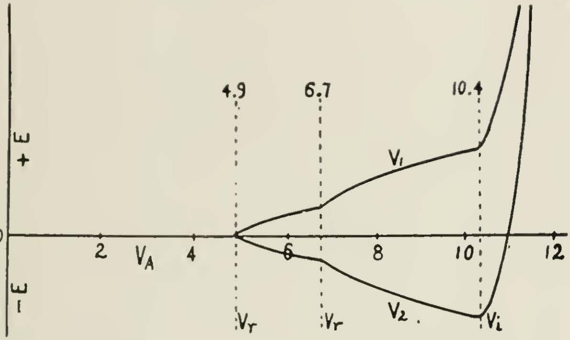


Fig. 4 Onset of ionization in mercury vapor at 10.4 volts (preceded by subsidiary effects at 4.9 volts and 6.7 volts; see footnote[6]). (Davis and Goucher)

we have hitherto imagined. So long as the energy of the electrons does not attain the resonance-potential of the gas, there is no reason to expect any novel effects; the collisions will be perfectly elastic, just as when the electrons were few. But when the atoms are thrown into the "excited state" by impacts, there will be occasional cases of an atom being struck twice by electrons in such quick succession that at the moment of the second blow, it is still in the excited state provoked by the first. Now, much less energy is required to ionize an atom when it is in the excited state than when it is normal; consequently when the electrons are so abundant that these pairs of

[6] The sudden upturn at 10.4 volts is the swift rise of current at the onset of ionization. The much less violent upturns at 4.9 and 6.7 volts are due to the electrons expelled from the metal parts of the apparatus by the radiation from the excited atoms. In the lower curve, by modifying the apparatus, the latter upturns are translated into downturns to distinguish them from the upturn which denotes ionization. This distinction was not realized until 1917, and in articles published between 1913 and 1917 the lowest resonance-potentials of gases are given as their ionizing potentials. Enormous improvements in the methods and technique of measuring these critical potentials, and recognizing of which kind they are, have been effected since then.

nearly-simultaneous collisions happen often, ionization will begin at the resonance-potential. *In a profuse electron-stream, the threshold potential for ionization is the lowest resonance-potential.* Another feature of the profuse discharge is, that when ionization does commence the current leaps up much more suddenly and violently than it does in the scanty discharge. This is because the electron-current is depressed at first by the space-charge effect, the repellence which the electrons crossing the gap exert against the electrons which are on the verge of starting; when positive ions first appear in the gap, they cancel the action of a great number of the traversing electrons, and the flow of electrons from the cathode to anode is immensely increased. I shall speak of this more extensively further on.

We return to the case of the feeble electron-stream. We have considered various things which an electron may do to a helium atom which it strikes—bouncing off harmlessly, or putting the atom into an excited state, or ionizing it; we have mentioned that each of the two latter actions commences at a critical value of energy, at the so-called *resonance* or *ionizing* potential, respectively; we have considered the effect of each of these actions upon the discharge. Have we listed all the possible interactions between atoms of matter and atoms of electricity, when electrons flow across helium? and if we knew all the resonance potentials and all the ionizing potentials[7] of helium, could we predict all the features of all electrical discharges in pure helium, whether in rarefied gas or in dense, whether the electron-stream be scanty or profuse? This is the general belief; whether justified, it is impossible to say. We evidently need another Maxwell or another Boltzmann, somebody exceedingly skilful in statistical reasoning, able to take the information we can provide about the possibility or the probability of various kinds of impacts, and deduce the state of affairs in the mixture of atoms, ions and electrons without getting hopelessly entangled in the frightful maze of equations into which his very first steps would certainly lead him. While awaiting him we have to content ourselves with our successes in interpreting the flow of electrons through very rarefied helium and the other noble gases and the metal vapors; and as for the discharges in denser gases

[7] I have simplified this passage somewhat so as not to retard the exposition. We know that an electron may "excite" a helium atom if its energy exceeds 19.75 volts, but this does not prove that it *must* do so; it is more reasonable to suppose that it has a certain chance of exciting the atom, zero when its energy is less than 19.75 volts, but greater than zero, and a certain function of its energy, when the latter exceeds 19.75 volts. We should know these functions for all the resonance-potentials and for the ionizing-potential; independent experiments to determine them have been performed, and no doubt will be multiplied.

we have to take the experimental data as we find them, and analyze them as best we may, not with too great an expectation of penetrating to the properties of the ultimate atoms; and yet, as we shall see, the analysis does in certain cases penetrate unexpectedly far.

3. The Flow of Electrons Across Dense Air, Nitrogen, Hydrogen and Similar Gases

The celebrated series of researches by Professor Townsend of Oxford and by his pupils, commenced in 1902 and continuing through the present, relate chiefly to such gases as hydrogen, nitrogen, oxygen and the familiar mixture of the last two which we breathe; and chiefly to these gases at densities much greater than we have hitherto considered—densities corresponding to such pressures as a thousandth or a hundredth of an atmosphere, therefore so great that an electron crossing over from a cathode to an anode a few centimetres away must collide with scores or hundreds of atoms. If a stream of electrons is poured into perfectly pure helium of such a density, we must not look for a sudden onset of ionization when the voltage between cathode and anode is raised just past 24.5, for the reason illustrated by those experiments of Compton and Benade—the electrons lose energy in all of their collisions, even the elastic ones, and arrive at the anode not with the full energy corresponding to its potential but with this energy diminished by what they lost on the way. In the familiar diatomic gases, the electrons lose much more energy in their ordinary collisions. I did not speak of these gases in the foregoing section, because experiments of the very same type as those which show the sharp distinction between elastic impacts and inelastic impacts in the noble gases and give the sharply-defined values of the resonance-potentials of these gases, yield comparatively vague and ill-defined data, when they are performed on hydrogen or air. In these gases, above all in active gases like oxygen or iodine, it is unlikely that any of the impacts, whether the electrons be moving rapidly or slowly, are truly elastic.[9]

[9] However, Foote and Mohler have obtained quite undeniable evidence of critical potentials, at which the loss of energy by the impinging electron is much greater than it is just below these potentials. The electron can transfer energy to (and receive energy from) a molecule in more different ways than to (from) an atom; such as by setting the molecule into rotation, or putting its constituent atoms into vibration relatively to one another. There is also the mysterious fact of "electron affinity"—an electron may adhere firmly to a non-ionized molecule. Numerous measurements of the rate at which electrons progress through a gas (a field of research which I have not space to consider here) indicate that at field strengths such as prevail in these experiments, adhesion of electrons to molecules is rare and transient.

Now if an electron on its way through the electric field from cathode to anode strikes atoms so often that it rarely has a chance to acquire more than say half a volt of energy from the field between one impact and the next, and if in each impact it loses most of the energy it has just acquired—if this condition prevails, we need not wonder that the voltage between the electrodes must be raised far beyond the ionizing-potential of the gas before there is the least sign of intensification of current.

In interpreting the experiments upon such gases and at such pressures as these last, it has been customary to make a more drastic assumption, the opposite extreme from the one which justified itself in dealing with rarefied helium; it is assumed that the electron surrenders at every impact all the energy which it has derived from the field since its last preceding impact. One may be inclined to make mental reservations in accepting so extreme an assumption, and it could almost certainly be advantageously modified; but as a tentative assumption it is successful enough to be legitimate. If it is true the electron can never build up a capital of energy step by step along its path; the only chances it will have to ionize will come at the ends of unusually long free flights.

Let us imagine a specific case *pour fixer les idées*: supposing the anode and the cathode to be parallel plates d apart, and representing the potential-difference between them by V and the field strength between them by X ($X = V/d$), we will set $d = 6$ cm., $V = 300$ volts, $X = 50$ volts/cm.; we will imagine that the interspace is filled with a gas having an ionizing-potential equal to 15 volts, and so dense that the average free path of an electron between collisions is one millimetre. I say that the *average* free path is 1 mm. long; if all the sixty free paths which the electron traverses in going from cathode to anode were equal, it would never acquire more than 5 volts of energy, and could never ionize an atom; but owing to the statistical distribution of free paths about the mean value, there will be a certain number out of the sixty which will be longer than three millimetres, and long enough, therefore, for the electron to acquire the 15 volts of energy which are necessary to ionize. In this case there will be $60\epsilon^{-2}$, about eight, of these long free paths. In each centimetre there will be $10\epsilon^{-2}$ of them. I will use the letter α' to designate this latter number, which is the number of atoms struck by the electron in each centimetre of its path, at moments at which it has energy enough to ionize an atom; α' is therefore the *number of chances to ionize* which the electron has *per centimetre*. The formula for α' is:

$$\alpha' = \frac{1}{\lambda}\epsilon^{-V_0/X\lambda} = C\epsilon^{-CV_0/X} = Bp\epsilon^{-BpV_0/X}, \tag{3}$$

in which V_0 represents the ionizing-potential of the gas; λ represents the mean free path of the electron; C, its reciprocal, is the number of collisions suffered by the electron in each centimetre of the path; and, since C is proportional to the pressure of the gas, it is replaced by Bp in the final formulation.[9]

It is already clear that the new assumption leads to a theory which requires a different language and a different set of ideas from those of the foregoing section. Not the ionizing-potential, but the number of ionizations performed by an electron in a centimetre of its path, is the quantity to be measured by experimental devices; not the voltage between the electrodes, but the field strength in the gas, is the factor which controls the phenomena.[10] In dealing with gases which are expected to conform to the theory, the appropriate procedure is to measure the number of molecules which an electron ionizes in a centimetre of its path, for all practical values of the field strength X and the density of the gas (or its pressure p) as independent variables. I will designate this number, following the usual practice, by α; if the theory is true it cannot be greater than α', it may be less. These quantities α and α' are statistical quantities, not like the ionizing-potential qualities of the individual atom or molecule, and this is a misfortune and disadvantage of the theory and of the experiments which it interprets; we are not, so to speak, in the presence of the ultimate atoms as before, we are one step removed from them, and this step a difficult one to take.

The measurement of α is effected by varying the distance d between anode and cathode, and determining the current as function of d. If N_0 electrons flow out of the cathode in a second, the ionization commences at the distance $d_0 = V/X$ from the cathode, and from that

[9] Since the number of free paths, out of a total number N_0, which exceed L in length is equal to $N_0 \exp(-L/\lambda)$; and since the potential-difference between the beginning and the end of the path of length L, if parallel to the field, is XL. It may be objected that the electrons bounce in all directions from their impacts, while the language of this paragraph implies that they are always moving exactly in the direction of the field. The rebuttal is, that if they do lose almost all of their energy in an impact, or all but an amount not much greater than the mean speed of thermal agitation, they will soon be swerved around completely into the direction of the field no matter in what direction they start out.

[10] The ionizing-potential determines the distance from the cathode at which ionization commences; this is equal to $d_0 = V_0/X$, and within this distance from the cathode there is no ionization and the theory does not apply; beyond this distance the ionization is controlled entirely by the field strength and by the number of inflowing electrons and the voltage between cathode and anode affects it only insofar as it affects these.

point onward the electron-stream increases exponentially, so that the current Ne arriving at the anode is

$$Ne = N_0 e \exp \alpha\,(d - d_0) \tag{4}$$

In Townsend's experiments the cathode was a zinc plate, the anode a film of silver spread upon a quartz plate; through little windows in the silver film a beam of ultraviolet light entered in from behind, crossed over the interspace and fell normally upon the zinc plate, and drove electrons out of it. The zinc plate was raised and lowered by a screw; the voltage-difference between it and the silver film was altered *pari passu* so that the field strength in the gas remained always the same. The current rose exponentially as the distance between the plates was increased, and thus α was determined. A typical set of data (relating to air at 4 mm. pressure, with a field strength of 700 volts/cm.) is plotted logarithmically in Fig. 5, the logarithm of the current as ordinate and the distance from anode to cathode as abscissa. The first few points lie close to a straight line, corresponding to an exponential curve such as equation (4) requires; the value deduced for α is 8.16. (The distance d_0 is about .35 mm. and has been ignored.) Of the divergence of the later points from the straight line I will speak further on.

Such an experiment shows that there *is* an α—that the theory is not at any rate in discord with the first obvious physical facts—and it gives the value of α for the existing values of X and p. Townsend performed many such measurements with different field strangths and different pressures, and so accumulated a large experimental material for determining α as function of the two variables p and X. To interpret these we will begin by making the tentative and temporary assumption that whenever a molecule is struck by an electron having energy enough to ionize it, it *is* ionized—that is, $\alpha' = \alpha$. Rewriting the equation (3) which expresses α' as function of p and X, we see that

$$\alpha'/p = B \exp(-BV_0 p/X) = f(X/p). \tag{5}$$

Therefore, if $\alpha' = \alpha$, the quotient of α by p is a function of X and p only in the combination X/p; or, whenever the pressure and the field strength are varied in the same proportion, the number of molecules ionized by an electron in a centimetre of its path varies proportionally with the pressure. I leave it to the reader to invent other ways of expressing (5) in words which illuminate various aspects of its physical meaning.

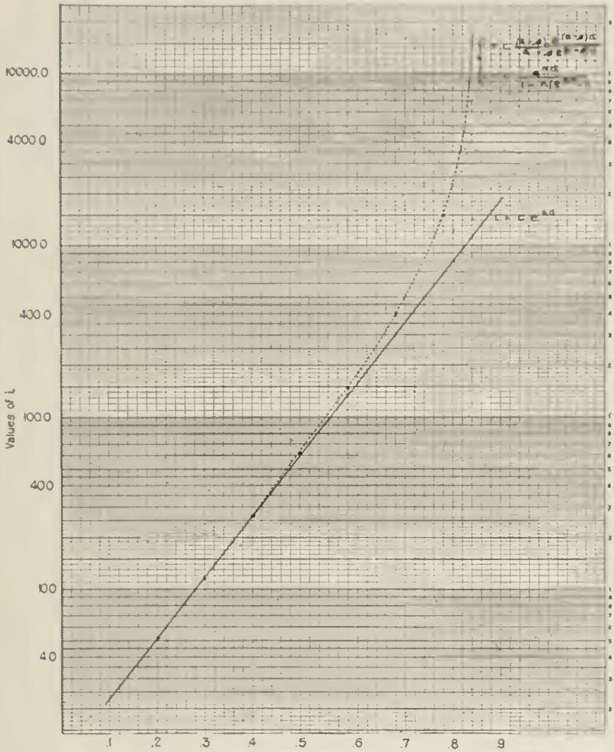


Fig. 5 Logarithmic plot of the currents across a gas (air) in which ionization by collision is occurring, for a constant field strength and various thicknesses of gas (Data from Townsend)

Experimentally, the test of (5) is made by dividing each one of Townsend's values of α by the pressure at which it was determined, and then plotting all these values of α/p versus the corresponding values of X/p. All the points for any one gas should lie on or close to a single curve, and within certain ranges of pressure and field strength they do; so far, good. The curve should be an exponential one, and within certain ranges of field strength and pressure it is; again, good. The next step is to calculate the values of B and V_0 which the curve imposes on the gas to which it relates. I quote the values of V_0, the ionizing-potential, which Townsend presents:

Air	N_2	H_2	CO_2	HCl	H_2O	A	He
25	27.6	26	23.3	16.5	22.4	17.3	12.3

When the first of these values were determined, no more direct way of measuring ionizing-potentials was known. Now that we have some values obtained by the direct methods sketched a few pages back, and fortified by indirect but very forcible evidence from spectroscopy, it is possible and quite important to test some of these. The values for argon and helium, although of the proper order of magnitude, are certainly too low. This is not in the least surprising, considering how many of the collisions between electrons and atoms must be perfectly elastic. It seems indeed rather mysterious that the current-voltage relation in either of these gases should have conformed closely enough to (4) to make it possible to define and measure α; but the electrons no doubt entered into many of the collisions with energy enough to put the atoms into excited states, if not to ionize them; and it is nearly always possible to take refuge in the assertion that the impurities may have been sufficient to distort the phenomena. As for the other gases in the list, all of them diatomic or triatomic, Townsend's values are too high—not very much too high, however; usually a matter of one-third to two-thirds.[11]

It appears therefore that the theory I have just developed is too simple, and must be amended. It seems natural to begin by dropping the tentative assumption that a molecule is ionized whenever it is hit by an electron having as much or more energy than is required to ionize it, and adopt instead the idea once already suggested in these pages, that it is sometimes but not always ionized by such a blow; that there is a certain *probability* of ionization by a blow from an electron having energy U, a probability which is zero when $U < V$ and is some yet-to-be-determined function of U when $U > V$. This would leave intact the conclusion that α/p should be a function of X/p, a conclusion which we have already found to be verified by experiment; but it would relieve us of the necessity of assuming that

[11] Townsend's values of B likewise correspond to values of the effective cross-section of the molecule, the quantity A of equation (2), which are of the same order of magnitude as the directly determined values of A.

that function is precisely the exponential function appearing in (5). Essentially the theory is reduced to this postulate: the number of molecules ionized by an electron in a centimetre of its path depends only upon the energy it acquires from the field in its free flight from one collision to the next. If in this form the theory still cannot give satisfaction, the next step will be to alter the original assumption that the electron comes practically to a dead stop in every collision. In dealing with the noble gases and the metal vapours, the facts about elastic collisions which I have already outlined prove that this assumption should not be made at all. It is clear that this is another problem for the future Boltzmann!

Meanwhile, one of the cardinal features of the Townsend experiments is the fact that they display the gradual advent of the transformation of the maintained currents which we have hitherto considered, into the self-maintaining discharges which are the familiar and the spectacular ones; and we now have to examine the agencies of this transformation.

4. The Discharge Begins to Contribute to the Electron-Stream Which Maintains It

Greatly though the current of primary electrons from the cathode to the anode may be amplified by the repeated ionizations which I have described, there is nothing in this process which suggests how the discharge may eventually be transformed into a self-maintaining one like the glow or the arc. The free electrons may ionize ever so abundantly, but as soon as the supply from the cathode is suspended by cutting off the heat or the light, the last electrons to be emitted will migrate off towards the anode, and whatever electrons they liberate will go along with them, leaving a stratum of gas devoid of electrons in their wake; and this stratum will widen outwards and keep on widening until it reaches the anode, and then the discharge will be ended. Something further must happen continually in the gas through which the electrons are flowing, something which continually supplies new free electrons to replace, not merely to supplement, the old ones which are absorbed into the anode and vanish from the scene.

We have already noticed one sort of event continually happening in such a gas as helium traversed by not-too-slow electrons, which might conceivably develop into a mechanism for maintaining the discharge; for, when an atom of the gas is put into the "excited state" by a blow from an electron, it later returns into its normal state, and in so returning it emits a quantum of radiant energy which may strike the cathode, and be absorbed by it, and cause another electron to leap out of the cathode and follow the first one. There are two other conceivable processes, which have the merit that they can not only be conceived but also witnessed in operation by themselves when the right conditions are provided. Positive ions flung violently against a metal plate drive electrons out of it, as can be shown by putting a positively-charged collector near the bombarded plate and noticing the current of negative charge which flows into it; and positive ions flowing rapidly across a gas ionize some of the atoms in it, as may be shown by sending a beam of such ions across the interspace between two metal plates, with a gentle crosswise field between them which sucks the freed electrons into the positive plate. The mechanism of the first process is not understood, except when the positive ions are so many and so swift that they make the metal hot enough to emit thermionic electrons, which does not happen in the cases we are now considering. The mechanism of the second process is only dimly understood, but it is clear enough that a positive ion driven against an atom is much less likely to ionize it, than an electron of equal energy would be.[12] Either of these two processes is very inefficient, at least at the comparatively low speeds with which positive ions move under the circumstances of these experiments; but they are probably efficient enough to do what is required of them. No doubt all three of them contribute to the discharge; but the relative proportions in which they act certainly differ very much from one sort of discharge to another, and will furnish research problems for years to come.

Returning to Fig. 5, we note once more that as the electrodes are moved farther and farther apart while the density of the gas and the field strength are held constant, the current at first rises exponentially (linearly in the logarithmic plot) as it should if the free electrons and only the free electrons ionize; but eventually it rises more rapidly and seems to be headed for an uncontrollable upward sweep. Townsend attributed this uprush to the tardy but potent participation of the positive ions, either ionizing the molecules of the gas by impact after the fashion of the negative ions, or driving electrons out of the cathode when they strike it, or both. Either assumption leads to

[12] If momentum is conserved in the impact between ion and atom, the ion must retain a large part of its kinetic energy after the collision, or else the struck atom must take a large part of it as kinetic energy of its own motion; it is not possible for the striking particle to spend nearly its entire energy merely in liberating an electron from the struck one. Conservation of momentum perhaps does not prevail on the atomic scale; but of all the principles of classical dynamics, it is the one which the reformers of physics most hesitate to lay violent hands upon.

an equation expressing the data equally well. If we adopt the former, and designate by β the number of molecules ionized by a positive ion in a centimetre of its path, and by N_0 the number of electrons supplied per second at the cathode, we get

$$N = \frac{N_0(\alpha-\beta)e^{(\alpha-\beta)d}}{\alpha-\beta e^{(\alpha-\beta)d}}. \tag{6}$$

Of course, β must be much smaller than α, or the positive ions would have made themselves felt earlier. Or if we adopt the latter idea, and designate by k the number of electrons expelled from the cathode (on the average) by each positive ion striking it, we arrive at the formula

$$N = \frac{N_0 e^{\alpha d}}{1-k(e^{\alpha d}-1)}. \tag{7}$$

Naturally k must be much smaller than unity for the same reason. In Fig. 5 the broken curve represents (6), with the values 8.16 and .0067 assigned to α and β; it also represents (7), with the values 8.16 and .00082 assigned to α and k.[13] (It was expected that the curves representing the two equations would be perceptibly apart on the scale of Fig. 5; but they were found to fall indistinguishably together.)

Evidently, therefore, the positive ions, weak and lethargic as they are in liberating electrons (one has only to compare β with α, or look at the value assigned to k in the last sentence!), can produce a notable addition to the current when the electrodes are far enough apart; and more than a notable addition, for when the distance d is raised to the value which makes the denominator of (6)—or of (7), whichever equation we are using—equal to zero, the value of N is infinite! Per-

[13] The derivations of (6) and (7) are as follows. Represent by $M(x)$ the number of electrons crossing the plane at x in unit time (the cathode being at $x=o$ and the anode at $x=d$); by $P(x)$ the number of positive ions crossing the plane at x in unit time; by N_0, the number of electrons *independently* supplied at the cathode per unit time, which is not necessarily equal to the value of M at $x=o$ (hence the notation); by i the current, or rather the current-density, as all these reasonings refer to a current-flow across unit area. We have

$$Me+Pe=i, \text{ hence}$$

making the assumption which leads to (6) we have

$$dM/dx = \alpha M+\beta P = (\alpha-\beta)M+\beta i/e$$

The boundary conditions are: $M=N_0$ at $x=o$ and $M=i/e$ at $x=d$. Integrating the equation and inserting these we get (6). Making the assumption which leads to (7) we have

$$dM/dx = \alpha M$$

The boundary conditions are: $M=N_0+k(i/e-M)$ or $(1+k)M=N_0+ki/e$ at $x=o$, and $M=i/e$ at $x=d$. Integrating the equation and inserting these we arrive at (7).

haps the best way to conceive of this is, that as the distance between the plates is increased toward that critical value of d, the value of N_0—which is the rate at which we have to supply electrons at the cathode, in order to keep a preassigned current flowing—diminishes continuously and approaches zero; so that eventually the current will keep itself going (and actually start itself) with the assistance of the occasional ions which are always appearing spontaneously in every gas, even though it be encased in an armor-plated shield. Of course, it is rather risky to predict just what is going to happen, when an equation which has been fixed up to represent a finite physical phenomenon over a certain range exhibits an infinite discontinuity at a point outside of that range. Usually, of course, the infinite value which the equation requires is modified into a finite one by the influence of some factor which was neglected when the equation was devised. In this case, however, the infinite discontinuity corresponds to a sudden catastrophic change. If an electrometer is shunted across the interspace between anode and cathode, its needle is forcibly jerked; if a telephone-receiver is connected in series with the interspace, it makes a clicking or a banging sound; if the gap is wide, so that the voltage just before the disruption is high, there is a brilliant flash, which may bear an uncomfortably strong resemblance to the lightning-bolt which is the cosmical prototype of all electric sparks.

What goes on after the critical moment of transition or transformation depends on many things; and not only on obvious features of the spark-gap, such as the kind and density of gas and the shape and size and material of the electrodes, but also on such things as the resistances and the inductances in series with the discharge, and the qualities of the source of electromotive force and its ability to satisfy the demands for current and voltage which the new discharge may make. Sometimes these demands are too extravagant for most laboratory sources or perhaps for any source to meet; probably this is why the spark between extended plane surfaces in dense air is as ephemeral as it is violent. But this does not always happen; in a sufficiently rarefied gas, the self-maintaining discharge which sets in after the transformation requires only a modest current and a practicable voltage, and supports itself with a few thousand volts applied across its terminals. The same thing occurs in a dense gas, if either of the electrodes is pointed or sharply curved, like a needle or a wire; the condition, more exactly, is that the radius of curvature of either electrode should be distinctly less than the least distance between the two. The transformation, however, is always very sudden, whether the new discharge be transient or permanent; and there are also sudden transitions from one sort of self-maintaining discharge to another, *e.g.*, from glow to arc or from one kind of glow to another when certain critical conditions are transgressed (critical conditions which may themselves depend on the battery and the circuit as well as the constants of the spark-gap). There are discontinuities of current and discontinuities of voltage at these transitions, and abrupt changes in the visible appearance in the discharge; and at each transformation there is a rearrangement of the distribution of space-charge in the gas. Hitherto we have encountered space-charge only in one or two of its simplest manifestations, retarding the flow of an electron-stream across a vacuum, and suddenly annulled when positive ions are mingled with the stream. Now we have to consider much subtler and more complicated cases, in which the space-charge varies rapidly in density and even in sign from one part of the gas to another, and the field and potential distributions are utterly distorted by it; and these distortions are essential to the life of the discharge. This distribution of space-charge is indeed dominant; and so I will write down some formulae which may be used to describe it.

5. Digression to Write Down Some Space-Charge Equations

The fundamental equation of the electrostatic field, known as Poisson's equation, is

$$\nabla^2 V = \frac{d^2V}{dx^2} + \frac{d^2V}{dy^2} + \frac{d^2V}{dz^2} = -4\pi\rho \tag{8}$$

in which V represents the electrostatic potential, and ρ the volume-density of electric charge.

We consider only the mathematically simplest case in which all variables are constant over each plane perpendicular to the x-axis, and so depend only on the coordinate x; as for example near the middle of an exceedingly wide tube with the x-axis lying along its axis. In this case Poisson's equation is

$$\frac{d^2V}{dx^2} = \frac{dX}{dx} = -4\pi\rho \tag{9}$$

in which X represents the potential-gradient, or field strength with sign reversed.[14] The value of X is determined at all points when the

[14] Field-gradient is therefore, proportional to space-charge with sign reversed, and *vice versa*. Positive field-gradient implies negative space-charge; negative field-

value of X at any one point and the values of ρ at all intermediate points are preassigned. Thus let X_0 represent the preassigned value of X at $x=0$, and X_d represent the value of X at $x=d$; we have

$$X_d = -4\pi \int_0^d \rho \, dx + X_0. \tag{10}$$

Consequently the P.D. between any two points is also determined; that between $x=0$ and $x=d$ is

$$V_d - V_0 = -4\pi \int_0^d dx \int_0^x \rho \, dx + X_0 d. \tag{11}$$

Now we introduce the further assumption that the electric charge is concentrated upon corpuscles (electrons or charged atoms) of one kind, of equal charge E and mass m, of which there are ndv in a very small volume dv at x; n is a function of x. Then

$$nE = \rho. \tag{12}$$

Assume finally that the corpuscles are moving with speed u, identical for all corpuscles having the same x-coordinate, but depending on x; represent the current-density by i; we have

$$nEu = i \tag{13}$$

and consequently

$$\rho = i/u. \tag{14}$$

Now consider the flow of current between two parallel planes, from one electrode at $x=0$ to the other at $x=d$. If the current is borne by corpuscles of one kind, and the assumption last made is true; and if we know the speed of the corpuscles at every point between the plates, and the field strength at some one point; then we can calculate the field strength everywhere between the plates, and the potential-difference between them.

The customary convention about the field strength is to assume it to be zero at the electrode from which the corpuscles start, so that $X_0=0$ in (11). Rewriting (11) to take account of (14), we have

$$V_d - V_0 = -4\pi i \int_0^d dx \int_0^x dx/u \tag{15}$$

as the general equation.

gradient implies positive space-charge, uniform field implies zero space-charge. It is instructive to examine mappings of field-distribution with this principle in mind, such mappings, for example, as those in Fig. 9. The uniform field in a current-carrying wire means that positive and negative charges are distributed everywhere in the metal with equal density—a conclusion one might forget, but for these more general cases.

If we suppose that the corpuscles acquire their speed u at the distance x in free flight from the electrode where they start, we have $\frac{1}{2}mu^2 = eV$, and

$$(V_d - V_0)^{3/2} = \frac{9\pi}{\sqrt{2}}\sqrt{\frac{m}{E}}\, id^2. \tag{16}$$

This is the equation adapted to electrons or other ions flowing across otherwise empty space.

If we suppose that the corpuscles have at each point a speed proportional to the field strength at that point, we have $u = \pm k\, dV/dx$, and

$$V_d - V_0 = \frac{2}{3}\sqrt{\frac{8\pi i\, d^3}{k}}. \tag{17}$$

This equation would be adapted to ions drifting in so dense a gas, or so weak a field, that they acquire *very little* energy from the field (in comparison with their average energy of thermal agitation in the gas) between one collision and the next, and lose it all at the next.[15]

If we conceive of ions which acquire *much* energy from the field between one collision and the next (much, that is, in comparison with their average energy of thermal agitation) and lose it all at the next collision, we have $u^2 = (\pi el/2m)\, dV/dx$ and

$$(V_d - V_0)^{3/2} = Cid^{5/2} \tag{18}$$

the constant C being equal to $\sqrt{m/El}$ multiplied by a certain numerical factor, and l standing for the mean distance traveled by the ion between one collision and the next.

The theory just given is too simple; it is an essential fact of the actual physical case that the ions emerge, at the surface of the electrode whence they start, with forward velocities which are distributed in some way or other about a mean value. These initial forward velocities, though often small compared with the velocities which the ions may acquire as they cross to the other electrode, are large enough so that all of the ions would shoot across the gap if the field strength were really zero at the emitting electrode and assisted them everywhere beyond it. In fact the space-charge creates a retarding field at the surface of the emitting electrode, and a potential minimum (if the ions are negative; a potential maximum, if the ions are positive) at a certain distance in front of it. Here, and not at the emitting electrode as we previously assumed, the field strength is zero. Equation (16) is often valid in practice, because this locus of zero field-strength is often very close to the emitting electrode. In fact, by

[15] As in electrical conduction in solid metals (cf. my preceding article).

raising the P.D. between the plates sufficiently, the locus of zero field can be driven back into coincidence with the emitting plate; beyond which stage, the "limitation of current by space-charge" ceases. But if the P.D. is sufficiently low the potential minimum (or maximum) is prominent and is remote from the electrode, and in these cases the equations we have just deduced are inapplicable.

It thus may readily happen that when we apply a certain potential to one electrode and a certain other potential to another electrode separated from the first one by gas or vacuum, we may find points between them where the potential is *not intermediate* between the potentials of the electrodes. This is a queer conclusion, to anybody accustomed to the flow of electricity in wires. But it is true, and must be kept in mind.

6. The Self-Maintaining Discharges

The *Arc* ought to be the easiest to understand among the self-maintaining discharges, in one respect at least; for it keeps its own cathode so intensely hot that thermionic electrons are supplied continuously in great abundance at the negative end of the discharge, and the theorist can begin his labors by trying to explain how and why this high temperature is maintained. Anything which tends to lower the temperature of the cathode, for instance by draining heat away from it, is very perilous to the arc. Stark uses various schemes for preventing the cathode from growing very hot, and they all killed the arc. This also explains why the arc is most difficult to kindle and most inclined to flicker out when formed between electrodes of a metal which conducts heat exceptionally well, and most durable when formed between electrodes of carbon, which is a comparatively poor conductor for heat. It probably explains why the arc has a harder time to keep itself alive in hydrogen, a gas of high thermal conductivity, than in air. While the gas in which the arc has its being and the anode to which it extends both influence the discharge, the high temperature of the cathode is cardinal.

The cathode is presumably kept hot by the rain of positive ions upon it, striking it with violence and yielding up their energy of motion to it; at least this is the obvious and plausible explanation. Now the arc is commonly and easily maintained in fairly dense gases, with a comparatively small potential-difference between widely-separated electrodes; and the energy which an ion can acquire from the field strength prevailing in it, in the short interval between two collisions with molecules, is so small that it cannot be made to account for the furious heat developed at the cathode when the ions finally strike it. Just before the ions arrive at the cathode they must be endowed with a kinetic energy which is very unusual (to say the least) in the middle of the discharge; and it is in fact observed that just in front of the cathode there is a sharp and sudden potential-fall, corresponding to a strong field extending but a little way outward from the electrode and then dying down into the weak field prevailing through the rest of the arc. This strong field picks up the ions which have meandered to its outward edge from the body of the discharge and hurls them against the cathode—not very forcibly, for the energy they receive from that potential-fall is not a great amount by ordinary standards, and most of the ions probably lose some of it in collisions on the way; but with much more energy than they would be likely to possess anywhere else in the arc.

This potential-fall immediately in front of the negative electrode, the *cathode-fall* of the arc, is measured by thrusting a probe or sounding-wire into the discharge as close as possible to the cathode (generally about a millimetre away), and determining the *P.D.* between it and the cathode. The probe is regarded with some distrust, as it raises in an acute form the old question as to how far the phenomena we observe in nature are distorted by the fact that we are observing them; the wire may alter the potential of the point where it is placed, or it may assume a potential entirely different from that of the environing gas; but the general tendency nowadays, I believe, is to accept its potential as a moderately reliable index of the potential which would exist at the point where it stands if it were not there.[16] The cathode-fall, as so measured, depends unfortunately on quite a number of things; the material of the cathode, the gas, the current. The gas is always mixed with a vapour of the electrode-material, particularly in the vicinity of the electrode; the only way to have a single pure gas is to enclose the whole system in a tube, evacuate the tube to the highest possible degree, and then heat it until the vapor-tension of the metal of which the cathode is made rises high enough for the vapor to sustain the arc. This is practicable with the more fusible metals; and with mercury, the arc generates heat enough to maintain the vapor-tension sufficiently high. In pure mercury-

[16] On this matter the experiments of Langmuir and Schottky, mentioned further along in this article, promise new knowledge. The probe automatically assumes such a potential that the net current-flow into it is nil; for example, if it is immersed in an ionized gas in which electrons and ionized atoms are roaming about, its eventual potential is such that equal numbers of particles of the two kinds strike and are absorbed in it per unit time. If the electrons are much more numerous or have a much higher average energy, or both, this potential may be several volts more negative than the potential at the same point before the probe was put in. The same may be said about the wall of the tube.

vapor, the cathode-fall assumes the value 4.9 volts which is the first resonance-potential of the mercury atom and therefore, as we have seen, is effectively the ionizing-potential of the free mercury atom when the electron-stream is as dense as it is in the arc. This suggests a delightfully simple theory of the whole process: the electrons stream from the cathode, they acquire 4.9 volts of energy from the cathode-fall, they ionize mercury atoms at the outward edge of the region of high field strength, the positive ions so created fall backward across the cathode-fall and strike the cathode, surrender their energy to it and so keep it hot, more electrons pour out, and so forth *ad infinitum*. It remains to be seen whether so simple a theory can be modified, by statistical considerations or otherwise, to explain the values of the cathode-fall in mixed and diatomic gases.

We do not know *a priori* what is the ratio of the number of electrons flowing outward across the cathode-fall in a second to the number of ions flowing inward. It might, however, be very great, and still the number of ions within the region of the cathode-fall at any instant could far surpass the number of electrons within it—the electron moves so much more rapidly than the ion, and has so much better a chance of crossing the region in one free flight without a collision. Even in hydrogen, in which the ions are the lightest of all ions, the electron current would have to be 350 times as great as the ion-current if the electrons just balanced the ions in unit volume. It is therefore legitimate to try out the assumption that the region of cathode-fall is a region of purely positive space-charge, in which some such equation as (16), (17), or (18) gives the current of positive ions as a function of the cathode-fall and the width of the region. K. T. Compton selected (18). Unfortunately the width of the cathode-fall region has not been measured, but he assumed it equal to the mean free path of an electron in the gas. The value which he thus calculated for the current of positive ions was about 1% of the observed total current; the remaining 99% consists of the electrons.

From the cathode region onward to the anode, the gas traversed by the arc is dazzlingly brilliant. In the long cylindrical tubes which enclose the mercury arcs so commonly seen in laboratories and studios, the vapor shines everywhere except near the ends with a cold and rather ghastly white light tinged with bluish-green. This is the positive column of the mercury arc. The potential-gradient along it is uniform, suggesting the flow of electricity down a wire; but here the resemblance stops, for when the current-density goes up the potential-gradient goes down. The curve of voltage versus current, which for a solid metal is as we all know an upward-slanting straight line, is for the arc a downward-slanting curve (Fig. 6). Such a curve is called a *characteristic*, and the arc is said to have a *negative characteristic*. Ionization goes on continually within the positive column, and ions of both signs can be drawn out by a crosswise field; but recombination of ions, a process which we have not considered, also goes on continually and maintains an equilibrium. Presumably it

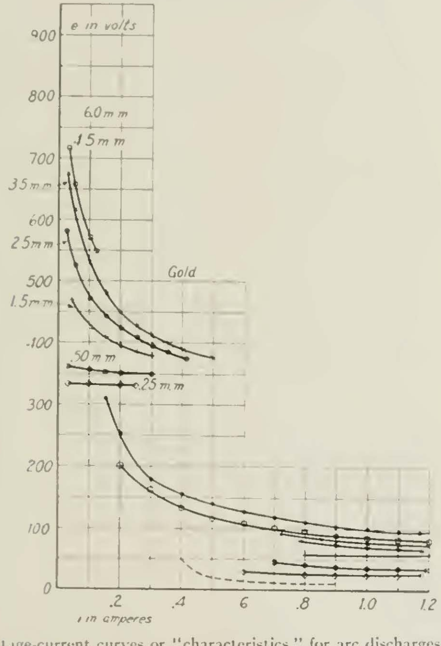


Fig. 6—Voltage-current curves or "characteristics," for arc discharges (below) and glow discharges (above) in air, between gold electrodes. The different curves correspond to different anode-cathode distances. (Ives, *Journal of the Franklin Institute*)

is the effect of the field strength on this equilibrium which causes the current-voltage curve to slant in what most people instinctively feel is the wrong way; but the theory of the equilibrium is not yet far advanced.

Langmuir and Schottky, working independently in Schenectady and in Germany, performed some very pretty experiments by thrusting negatively-charged wires or plates into the positive column. These wires and plates surrounded themselves with dark sheaths, the thickness of which increased as the potential of the metal was made more and more highly negative. The explanation is, that the electrons in the positive column cannot approach the intruded wire, being driven back by the adverse field; the dark sheath is the region from which they are excluded, and across it the positive ions advance to the wire through a field controlled by their space-charge. The equation selected by Langmuir to represent the relation between the thickness of the sheath, the voltage across it, and the current of positive ions into it, is (16). As the sheath is visible and its thickness can be measured, as well as the other quantities, the relation can be tested. This was done by Schottky; the result was satisfactory. When the intruded electrode is a wire, the sheath is cylindrical, and expands as the voltage of the wire is made more negative. As the area of the outer boundary of the sheath is increased by this expansion, more ions from the positive column touch it and are sucked in, and the density of flow of positive ions in the column can be determined. By lowering the potential of the wire gradually so that the electrons can reach it, first the fastest and then the slower ones, the velocity-distribution of the electrons in the column can be ascertained. Their average energy depends on the density of the mercury vapour, and may amount to several volts.

Beyond the positive column lies the anode, itself preceded by a sharp and sudden potential rise. The electrons are flung against it with some force, and it grows and remains very hot; usually, in fact, hotter than the cathode. This high temperature does not seem to be essential to the continuance of the discharge, for the anode can be cooled without killing the arc; yet it seems strange that a quality so regularly found should be without influence upon the discharge. One must beware of underestimating the influence of the anode; when an arc is formed in air between two electrodes of different materials, it behaves like an arc formed between two electrodes of the same material as the anode, *not* the cathode!

The so-called *low-voltage arc*, although not a self-maintaining discharge, merits at least a paragraph. A dense electron-stream poured into a monatomic gas from an independently-heated wire, and accelerated by a P.D. surpassing the resonance-potential of the gas, may ionize it so intensely that there is a sudden transformation into a luminous arc-like discharge. This is a sort of "assisted" arc, its cathode being kept warm for its benefit by outside agencies. Its history is a long and interesting chapter of contemporary physics, whereof the end is not yet. The most remarkable feature of this arc

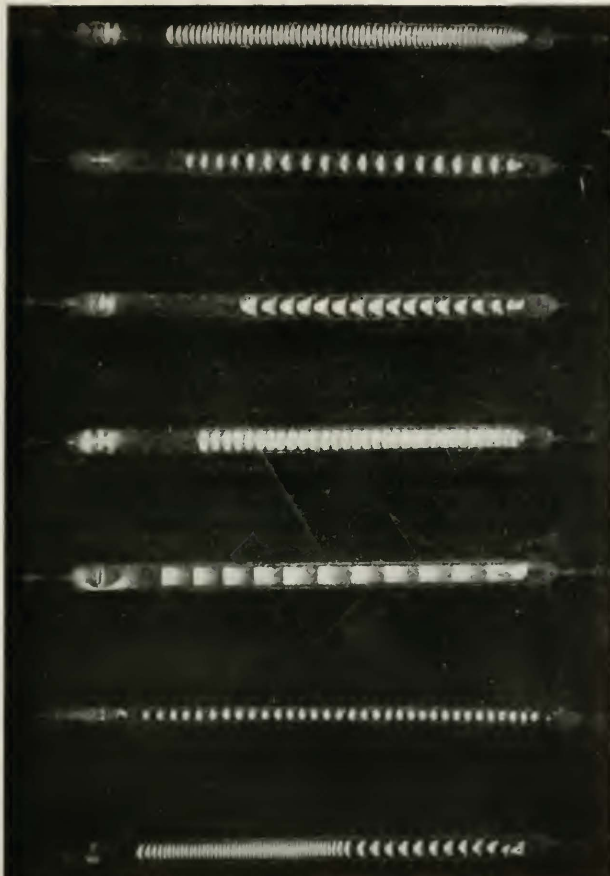


Fig. 7—Photographs of the glow-discharge in a long narrow cylinder, showing chiefly the subdivision of the positive column into striations. (De la Rue and Muller, *Philosophical Transactions of the Royal Society*)

is that it can survive even if the voltage between anode and cathode is far below the resonance-potential of the atoms of the gas, which seems impossible. A year ago it seemed that this effect could always be ascribed to high-voltage high-frequency oscillations generated in the arc. This explanation was presently confirmed in some cases and disqualified in others, and now it appears that when there are no oscillations an astonishingly strong potential-maximum develops within the ionized gas. Potential-maximum and oscillations alike are probably to be regarded as manifestations of space-charge.

The *Glow in a rarefied gas* is a magnificent sight when the gas is rarefied to the proper degree, not too little and not too much; divided into luminous clouds of diverse brightnesses and diverse colors, recalling Tennyson's "fluid haze of light," yet almost rigidly fixed in their distances and their proportions, it is one of the most theatrical spectacles in the repertoire of the physical laboratory. The grand divisions of the completely-developed discharge are four in number, two relatively dim and two bright; beginning from the cathode end,

Fig. 8—The Crookes dark space between the cathode (thin line at left) and the negative glow. See footnote [12]. (Aston, *Proceedings of the Royal Society*)

they are the Crookes dark space, the negative glow, the Faraday dark space, and the positive column. Additional gradations of color and brightness can often be seen very close to the cathode and very close to the anode. Photographs of the glow which give anything approaching a true idea of its appearance to the eye are hard to find. I reproduce in Fig. 7 some photographs taken nearly fifty years ago by de la Rue, which have reappeared in many a text; they show chiefly the striking flocculent cloudlets into which the positive column sometimes divides itself. In Fig. 9 there are two sketches made by Graham.

The Crookes dark space (or cathode dark space, or Hittorf dark space as it is called in Germany) extends from the cathode to the boundary of the bright luminous cloud which is the negative glow. The boundary is generally so well-defined and distinct that an observer finds it easy to judge when a sounding-wire just touches it, or the cross-hair of a telescope coincides with its image; "in the case of oxygen," Aston said, "the sharpness was simply amazing; even with so large a dark space as 3 cm., the sighter could be set (to the boundary) as accurately as to the cathode itself, *i.e.*, to about 0.01 mm." I reproduce some of Aston's photographs in Figure 8, although he says that for reasons of perspective the boundary of the negative glow appears more diffuse than it really is.[17] The electric field strength within the Crookes dark space is greater, often very much greater, than in any of the other divisions of the discharge; almost the whole of the voltage-rise from cathode to anode is comprised within it, and the remainder, although spread across all the brilliant parts of the glow, is inconsiderable unless the tube is made unusually long. The behavior of the dark space when the current through the tube is varied (by varying a resistance in series with the tube) is curious and instructive. If the current is small and the cathode large (a wide metal plate) the negative glow overarches a small portion of the cathode surface, lying above it like a canopy with the thin dark sheath beneath it. When the current is increased the canopy spreads out, keeping its distance from the metal surface unaltered, but increasing its area proportionally to the current; the thickness of the Crookes dark space and the current-*density* across it remain unchanged. If the experimenter continues to increase the current after the cathode is completely overhung by the glow, the dark space thickens steadily, and the current-density across it rises.

The changes in the voltage across the Crookes dark space which accompany these changes in area and thickness are very important. The voltage is measured with a sounding-wire, like the cathode-fall in the arc; but since the boundary of the dark space is so sharply marked, the experimenter can set the sounding-wire accurately to it instead of merely as close as possible to the cathode. So long as the

[17] Adjacent to the cathode a thin perfectly dark stratum can be distinguished (especially in the picture on the right). The P.D. across this thin black space is, as nearly as it can be guessed from the width of the space, of about the magnitude of the ionizing-potential of the gas. In fact Aston estimated it for helium (to which the pictures refer) as 30 volts, a good anticipation of the value 24.5 assigned years later to the ionizing potential. It seems therefore that the outer edge of the very dark space is at the level where the electrons coming from the cathode first acquire energy enough to ionize.

negative glow does not overarch the whole cathode, and the thickness and current-density of the dark space keep their fixed minimum values, the voltage across it remains constant likewise. This is the *normal cathode-fall* of the glow. It is an even more thoroughgoing constant than the thickness or the current-density of the dark space, for these vary with the pressure of the gas (the dark space shrinks both in depth and in sidewise extension, if the current is kept constant while the gas is made denser) while the normal cathode-fall is immune to changes in pressure. It depends both on the gas and on the material of the cathode; the recorded values extend from about 60 volts (alkali-metal cathodes) to about 100 volts. Attempts have been made to correlate it with the thermionic work-function of the cathode metal, and there is no doubt that high values of the one tend to go with high values of the other, and low with low. When the cathode is entirely overspread by the negative glow and the dark space begins to thicken, the voltage across it rises rapidly; the cathode-fall is said to become *anomalous*, and may ascend to thousands of volts.

Almost the whole of the voltage-rise from cathode to anode, as I have stated, is generally comprised in the cathode-fall; the remainder, although spread over all of the brilliant divisions of the discharge, is inconsiderable unless the tube is unusually long. The field strength in the Crookes dark space is also much greater than anywhere else in the glow. This is illustrated by the two curves in Fig. 9, representing the field strength in the discharges sketched above them. (For the region of the Crookes dark space, however, the curves are defective.) In the luminous clouds the electric force is feeble, and they in fact are not essential to the current-flow; if the anode is pushed inwards towards the cathode, it simply swallows them up in succession without interfering with the current; but the moment it invades the Crookes dark space, the discharge ceases unless the electromotive force in the circuit is hastily pushed up. The mechanism which keeps the glow alive lies concealed in the dark space.

One naturally tries to invent a mechanism resembling the one suggested for the arc: the cathode-fall serves to give energy to the electrons emerging from the cathode, so that they ionize molecules at the edge of the negative glow; and the ions fall against the cathode with energy enough to drive out new electrons. But the details are more difficult to explain. The cathode-fall gives much more energy to the electrons than they need to ionize any known molecule, so that apparently its high value is what the ions require to give them enough energy to extract electrons from the cathode. We can hardly argue that the electrons are thermionic electrons; the cathode does not grow hot enough; if it does, the cathode-fall suddenly collapses, and the glow is liable to turn into an arc. Expulsion of electrons from cold metals by ions striking them has been separately studied, but not sufficiently.

On the other hand, there is good evidence that the Crookes dark space, like those dark sheaths scooped out in the positive column

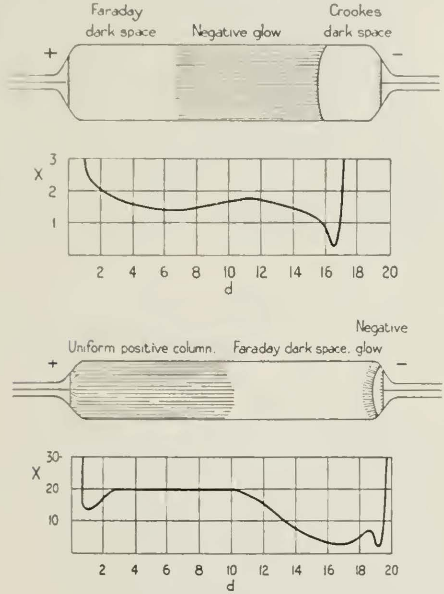


Fig. 9—Sketches of the glow in rarefied nitrogen at two pressures (the higher below) with curves showing the trend of field strength along the discharge. (Graham, *Wiedemanns Annalen*)

of the mercury arc by intruding a negatively-charged wire, is a region of predominantly positive space-charge, in which positive ions advance towards the cathode in a manner controlled by some such equation as (16) or (17). For example, Gunther-Schulze proposed (16) to describe the state of affairs in the Crookes dark space in the condition of normal cathode-fall; that is, he assumed that the ions fall unimpeded from the edge of the negative glow to the cathode surface. No doubt this assumption is too extreme, yet it leads to unexpectedly good agreements with experiment. Thus when the thickness of the Crookes dark space is altered (by altering the pressure of the gas) leaving the voltage across it constant, the current-density varies inversely as the square of the thickness, as it should by (16). And when Gunther-Schulze calculated the thickness of the dark space from (16), using the observed values of cathode-fall and current for six gases and two kinds of metal, and substituting the mass of the molecule of the gas for the coefficient m in that equation, the values he obtained agreed fairly well (within 10%) with the observed thicknesses. Long before, J. J. Thomson had proposed (17), and Aston tested it by a series of experiments on four gases, in the condition of strong anomalous cathode-fall. As k of that equation should be inversely proportional to the pressure p of the gas, the product id^3V^{-2} (V standing for the cathode-fall) should be constant at constant pressure, and the product $id^3V^{-2}p$ should be constant under all circumstances. These conclusions were fairly well confirmed for large current-densities.

Several attempts to test the theory by actually determining the potential-distribution in the Crookes dark space were made with sounding-wires and by other methods; but they have all been superseded, wherever possible, by the beautiful method founded on the discovery that certain spectrum lines are split into components when the molecule emitting them is floating in an intense electric field, and the separation of the components is proportional to the strength of the field. This was established by Stark who applied a strong controllable electric field to radiating atoms, and by LoSurdo who examined the lines emitted by molecules rushing through the strong field in the Crookes dark space, in the condition of anomalous cathode-fall. Now that the effect has been thoroughly studied it is legitimate to turn the experiments around and use the appearance of the split lines as an index of the field strength in the place where they are emitted. Brose in Germany and Foster at Yale did this. In the photographs (Fig. 10, 11) we see the components merged together at the top, which is at the edge of the negative glow, where the field is very small; thence they diverge to a maximum separation, and finally approach one another very slightly before reaching the bottom, which is at the cathode surface.[18] This shows that the net space-charge in the Crooke

[18] The displacements of certain components are not rigorously proportional to the field, and sometimes entirely new lines make their appearance at hitherto unoccupied places when a strong field is applied. Both of these anomalies can be detected in the pictures. For the original plate from which Fig. 11 was made I am indebted to Dr. Foster.

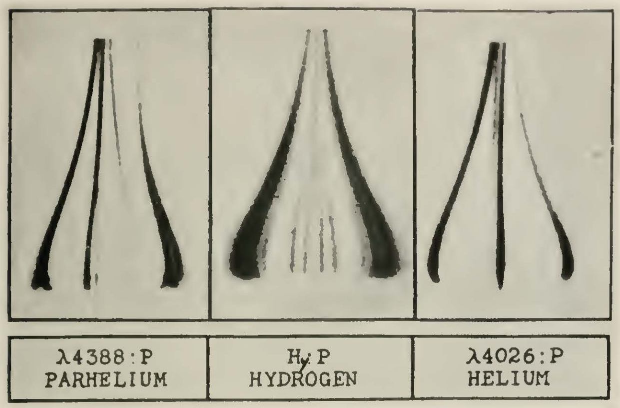


Fig. 10—Spectrum lines subdivided and spread out in the Crookes dark space by the strong and variable field. See footnote [18]. (J. S. Foster, *Physical Review*)

Fig. 11—A group of lines near λ4388 (parhelium spectrum) resolved and spread out in the Crookes dark space. See footnote [18].

dark space is positive from the edge of the negative glow almost but not quite to the cathode; there is a thin region just above the cathode where there is more negative charge than positive. This is splendid material for the theorist, and it is deplorable that the method cannot be applied except when the cathode-fall is anomalous and exceedingly large.

When a narrow straight hole is pierced in the cathode, the positive ions making for it shoot clear through, and can be manipulated in a chamber provided behind the cathode. In particular the ratios of their charges to their masses can be measured, and thence their masses can be inferred. This is Thomson's "positive-ray analysis," which Aston developed into the most generally available of all methods for analyzing elements into their isotopes. If the density of the gas is so far reduced that the Crookes dark space extends to the anode, the electrons can be studied in the same way and their charge-mass ratio determined. Hence the mass of the electron can be deduced, and its dependence upon the speed of the electron ascertained, yielding precious evidence in support of the special or restricted theory of relativity. These are among the simple phenomena which I mentioned at the beginning of this article, in which the properties of the ultimate atoms of electricity and matter are revealed.

The positive column, which is the brilliant, colorful and conspicuous part of the glow, resembles in some ways the positive column of the mercury arc. In it the potential-gradient decreases with increasing current, and the characteristic of the glow is negative (Fig. 6). Often the positive column subdivides itself into a regular procession of cloudlets or *striations*, all just alike and equally spaced (Fig. 7). The potential-difference between two consecutive striations has the same value all along the procession, and everyone feels instinctively that it ought to be the ionizing-potential or the resonance-potential of the gas; but this is evidently too simple an interpretation for the general case, although striations at potential-intervals of 4.9 volts have been realized in mercury vapor. Generally, if not always, the striations appear when the gas is contaminated with a small admixture of another. In this fact the key to the problem of their origin probably lies.

The *Glow in a dense gas* (as dense as the atmosphere, or more so) is visible only when the surface of either or both electrodes is curved, with a radius of curvature smaller than the minimum distance between the two. In these circumstances the field strength varies very greatly from one point to another of the interspace, at least before the space-charges begin to distort the field, and presumably afterwards as well;

it attains values just in front of the curved electrode (or electrodes, if both are curved) so great that if they prevailed over an equal interspace between flat electrodes they would instantly provoke an explosive spark. In some cases the glow in a dense gas resembles a very con-

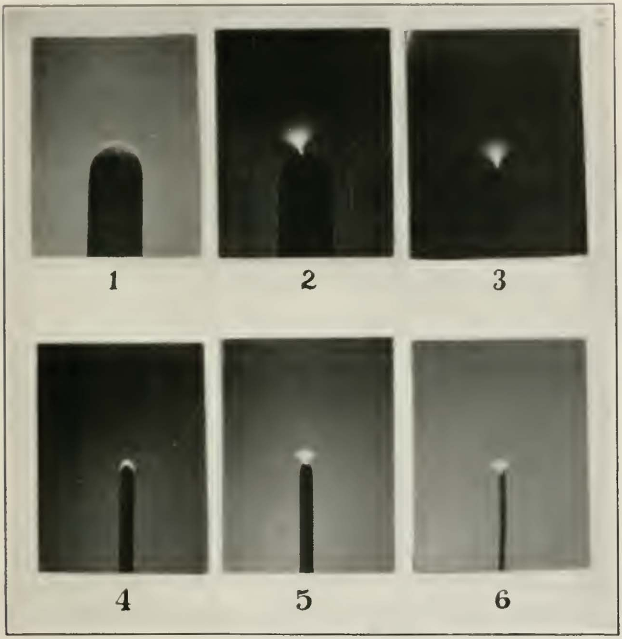


Fig. 12—The glow in air at atmospheric pressures, near a curved electrode (the other electrode is a plate beyond the top of the picture). In 1, 4 the curved electrode is the anode; in 2, 3, 5, 6 it is the cathode. (J. Zeleny, *Physical Review*)

tracted and reduced copy of portions of the glow in a rarefied gas. Thus in the photographs (Figs. 12, 13) of the luminosity surrounding a very curved cathode, it is possible to discern two dark spaces and two bright ones, the first dark space lying just outside the cathode, the last bright region fading off into the darkness which extends away towards the flat anode (far above and out of the picture). In the pictures of the glow surrounding a very curved anode, we see only a luminous sheath spread over the metal surface (Fig. 12).[13] Mathematically the simplest case (at least before the space-charge begins to affect the field) is realized by a slender cylindrical wire stretched along the axis of a

Fig. 13—Magnification of one of the pictures in Fig. 11. (The lowest bright spot is a reflection in the cathode surface)

much wider hollow cylinder, the wall of which may be imagined to recede to infinity in the limiting case. In this case the glow bears the euphonious name of *corona*, and has been intensively studied because it wastes the power transmitted over high-tension lines.

[13] I am indebted to Professor J. Zeleny for plates from which these figures were made.

Often there is a luminous cylindrical sheath encasing the wire, and from the boundary of the sheath outwards to the outer cylinder the gas is dark. It is customary to assume that the dark region, like the other dark spaces we have considered, is traversed by a procession of ions of one sign, positive or negative as the case may be, moving at a speed proportional to the field and controlled by their own space-charge according to the equation in cylindrical coordinates corresponding to (17); and the experiments support this assumption to a certain extent.

I must use my last paragraph to erase the impression—inevitably to be given by an account so short as this, in which the understood phenomena must be stressed and the mysterious ones passed over—that the flow of electricity through gases is to be set down in minds and books as a perfected science, organized, interpreted and finished. Quite the contrary! there are as many obscure and mysterious things in this field of physics as there are in any other which has been explored with as much diligence. Its remarkable feature is not that most or many of the phenomena in it have been perfectly explained; but rather, that for those few which have been explained, the explanations are very simple and elegant; they are based on a few fundamental assumptions about atoms and electrons which are not difficult to adopt, for they are not merely plausible but actually *demonstrable.* Perhaps as time goes on all the phenomena will be explained from these same assumptions. There will be experimenters who modify the apparatus and the circumstances of past experiments so that all of the avoidable complications are avoided and the phenomena are simplified into lucid illustrations of the fundamental principles; and there will be theorists, who take the complicated phenomena as they are delivered over to us, and extend the power of mathematical analysis until it overcomes them. They may find it necessary to make other and further assumptions, beyond those we have introduced; at present it is commonly felt that ours may be sufficient. Whether posterity will agree with us in this, must be left for posterity to decide.

Carrier Telephony on High Voltage Power Lines

By W. V. WOLFE

Introduction

THE use of power from hydro-electric generating stations and central steam plants has increased until single companies serve a territory of many thousands of square miles and the problem of coordinating the distributing centers with the generating stations has steadily increased in complexity.

One of the essentials of this coordination is obviously an adequate system of communication and until the recent advent of high frequency telephony, this service was secured over privately owned telephone lines and over lines of public service telephone companies.

The advent of the power line carrier telephone system now offers a highly reliable and satisfactory means of communication in connection with the operation of power systems. This equipment has been designed to employ the power conductors as the transmission medium and to provide service as reliable as the power lines themselves with a low initial cost, a small maintenance charge, increased safety for the operating personnel and transmission comparable in quality and freedom from noise with that obtained on high grade commercial toll circuits.

Preliminary Problems

In proceeding with the development of the Western Electric Power Line Carrier Telephone System three major problems were encountered. It was first necessary to learn from field tests and close contact with power companies the characteristics of power lines and associated apparatus at high frequencies and the operating requirements for such a telephone system; second, it was necessary to develop a safe and efficient method for coupling the carrier apparatus to the power conductors and third, to select and develop circuits and equipment suited to this service.

The superiority of the full-metallic over the ground return high frequency circuit was easily established by comparative measurements of attenuation, noise and interference, and therefore the experimental work was largely confined to the former circuit.

High Frequency Attenuation of Power Lines

Since the measurement of the attenuation of a circuit ordinarily requires that the circuit be terminated in its surge impedance[1] to avoid reflection effects, the first step in determining the attenuation of the power line was to measure its surge impedance. After considering several methods for measuring this impedance, a substitution

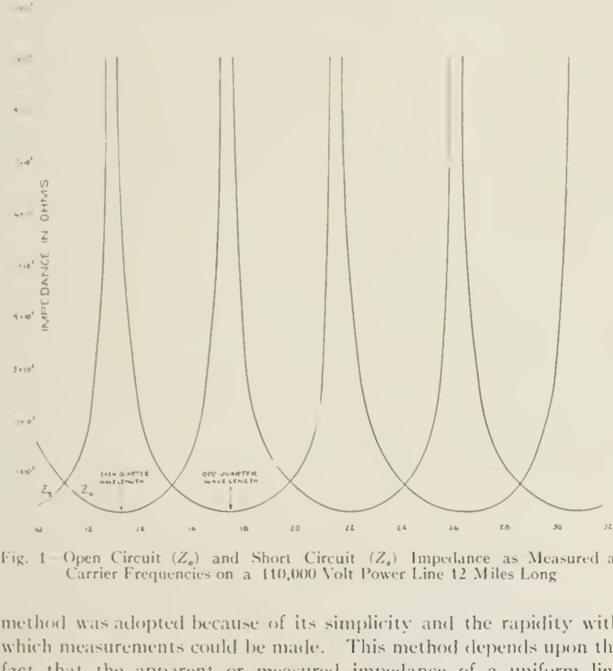


Fig. 1—Open Circuit (Z_o) and Short Circuit (Z_s) Impedance as Measured at Carrier Frequencies on a 110,000 Volt Power Line 12 Miles Long

method was adopted because of its simplicity and the rapidity with which measurements could be made. This method depends upon the fact that the apparent or measured impedance of a uniform line terminated in its surge impedance is equal to that surge impedance and it consists in terminating the line in a known resistance and determining the value of current supplied to the line by an oscillator

[1] Surge or characteristic impedance may be defined as the measured impedance of a uniform line of infinite length or in the case of a finite line it may be expressed mathematically as $Z = \sqrt{Z_{\text{open circuited}} \times Z_{\text{short circuited}}}$

and then substituting for the line a non-inductive resistance until the same value of current is drawn from the oscillator. In employing this method for determining the surge impedance it was assumed that the oscillator output was constant, and that the phase angle of the surge impedance was small.

A study of the curves on Fig. 1 shows that the apparent impedance of the line will change with the impedance in which the line is terminated in different ways, depending upon the frequency used. (1) If

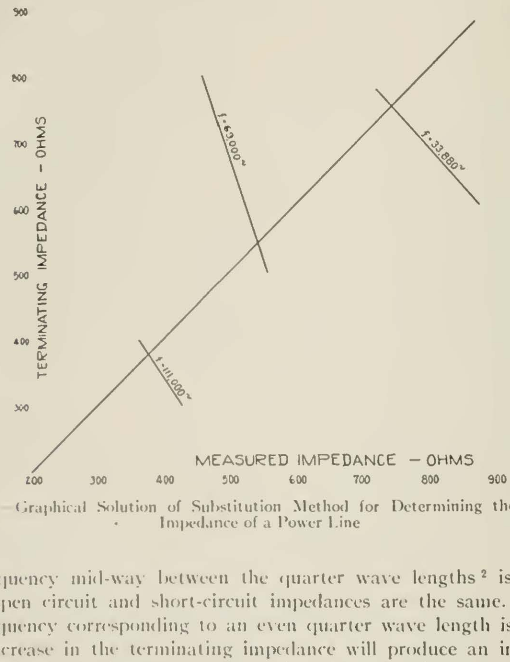


Fig. 2—Graphical Solution of Substitution Method for Determining the Surge Impedance of a Power Line

a frequency mid-way between the quarter wave lengths [2] is used, the open circuit and short-circuit impedances are the same. (2) If a frequency corresponding to an even quarter wave length is used, an increase in the terminating impedance will produce an increase in the apparent impedance of the line. (3) If a frequency corresponding to an odd quarter wave length is used, an increase in the terminating impedance will produce a decrease in the apparent imped-

[2] Whenever the length of the line becomes equal to, or some multiple of, one quarter of the length of the electric wave of the corresponding frequency, it is referred to as a quarter wave length frequency, or, for short, a quarter wave length.

ance of the line. If the apparent impedance of the line is plotted against the terminating impedance, in (1) the curve will be horizontal; in (2) the curve will have a positive slope approaching 45° and in (3) the curve will have a negative slope of approximately 45°. Each of these curves will intersect a 45° line drawn through the origin at a point where the terminal impedance is equal to the surge impedance of the line. This intersection can be determined with the

Fig. 3 Frequency vs. Attenuation and Frequency vs. Surge Impedance as Measured on the Tallulah Falls-Gainesville 110,000 Volt Power Line

greatest ease and accuracy when the curve crosses the 45° line at right angles or under condition (3), that is, when the determination is made at a frequency corresponding to an odd quarter wave length. To determine the surge impedance at a given frequency all that was necessary was to terminate the line at the distant end in an impedance which it was anticipated would be just below the surge impedance and measure by the substitution method the apparent impedance of the line, and then to terminate the line at the distant end in an impedance which would just exceed the surge impedance and determine the corresponding apparent impedance. The intersection of a straight line through these points with the 45° line determined the correct terminating impedance. In Fig. 2 is shown a determination of the characteristic impedance of the Tallulah Falls-Gainesville line of the Georgia Railway and Power Company at three different frequencies.

The attenuation of the line was then measured by terminating it in its characteristic impedance and measuring the current in to the line and current out of the line.[3] The results of the attenuation measurements made on the Tallulah Falls-Gainesville line are shown on Fig. 3. The irregularities in the attenuation shown by the lower

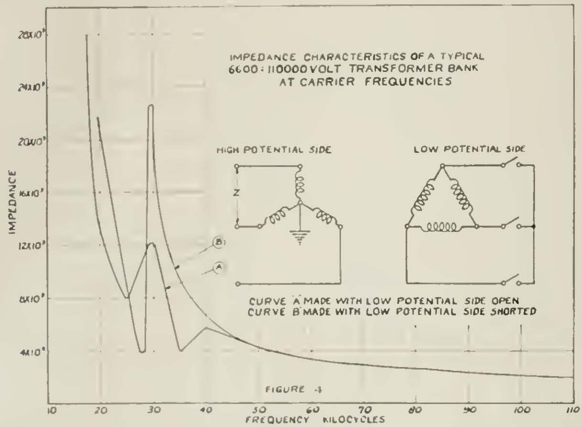


Fig. 4—Impedance Characteristics at Carrier Frequencies of a Typical 6600:110000 Volt Transformer Bank

curve are probably caused by the error in assuming that the phase angle of the surge impedance was small and that the surge impedance was a straight line function of frequency. From these and other data it was evident that for frequencies as high as 150 K.C. the attenuation is not excessive.

High Frequency Characteristics of Power Transformers

In order to determine the effect of power transformers on the use of the power line as a transmission medium for high frequency currents,

[3] Attenuation expressed in transmission units is equal to 20 $\log_{10} \frac{I_1}{I_2}$ where I_1 is the current into the network and I_2 is the current received from the network and measured in a circuit whose impedance corresponds to the characteristic impedance of the network.

the impedance of typical transformer banks was measured. In Fig. 4 is shown the impedance versus frequency characteristic of a three phase, 110,000 6600 V., 12,000 K.V.A. transformer bank connected "star" on the high side with the neutral grounded and "delta" on the low side. As shown by the diagram, these measurements were made between phases on the high side with the low side open circuited and short circuited. The coincidence of these curves for frequencies above 50 K.C. indicates that at these frequencies the dominant characteristic is the distributed capacity of the high winding and the impedance is probably unaffected by changes on the low potential side of the transformer. Below 50 K.C., however, the impedance changes rapidly both with frequency and with the low potential termination.

A study of Figs. 3 and 4 and other data shows that the desirable frequency range in which to operate a power line carrier telephone circuit is that from 50 K.C. to 150 K.C. In this range the attenuation is not excessive, it is very little affected by the associated power apparatus, and it is independent of the conditions on the low potential power circuits. The curve shown in Fig. 3 indicates that, contrary to the common belief, the attenuation in this range is a relatively smooth function of frequency. This conclusion is supported by the fact that in the various installations of power line carrier telephone equipment which have been made since the attenuation measurements on Fig. 3 were obtained, no power lines have been encountered where the attenuation was a critical function of frequency. Another important argument for the selection of this frequency range lies in the fact that it is well above the range employed for multiplex telephony on commercial telephone systems and therefore precludes any interference with such systems.

Coupling Between Carrier Equipment and Power Line

Probably the most difficult problem to solve was that of providing a satisfactory method for connecting the carrier equipment to the power line. The use of power transformers has not been found practicable for if frequencies low enough to be efficiently transformed were employed, the attenuation of the circuit would be a function of the conditions in the distributing network and a change in the number or arrangement of transformers would result in an appreciable change in the attenuation. Such a method of coupling to the power line would also have the objection that communication would not be possible when the power transformers were disconnected from the line.

Since it did not seem practicable to develop a carrier frequency transformer suitable for connecting between phases of a high voltage power line it was decided to couple to the power line by means of capacity. Two general types of condensers are possible, first, a concentrated capacity condenser and second, a distributed capacity condenser. A concentrated capacity condenser suitable for direct connection to a high voltage power line was not available, but its development has been successfully undertaken by the Ohio Brass Co.

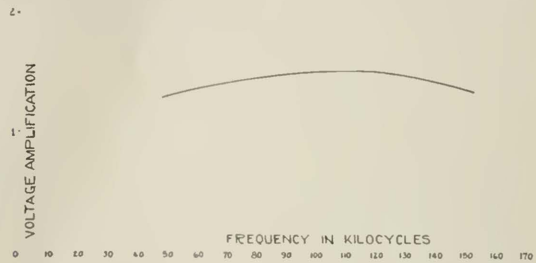


Fig. 5—Voltage Amplification Characteristic of High Frequency Transformer

The distributed capacity was obtained by suspending a wire parallel to the power conductor and employing this wire as one plate of the condenser and the conductor as the other plate. Both of these methods of connecting to the power line have been developed and are described later.

Design of the Carrier Equipment

Although the "carrier suppressed" system has many advantages over the "carrier transmitted" system, the difficulty of securing filters suitable for suppressing the unwanted products of the modulation prevented the use of the carrier suppressed system.

Several general characteristics of the electrical and mechanical design of this carrier equipment are worthy of note. The various stages of vacuum tubes in both the transmitting and receiving circuits are coupled by transformers. These transformers are closed iron core coils using the standard core employed for audio-frequency transformers. Fig. 5 shows the characteristic of one of these transformers, and it is evident from this figure that the variation in amplification from 50 K.C. to 150 K.C. is only a fraction of a transmission unit.

Although the frequencies employed by this equipment are fairly high, it was practicable to mount all of the apparatus on standard steel relay rack plates. In order to minimize the maintenance on this equipment no "C" batteries have been employed, the grid potentials

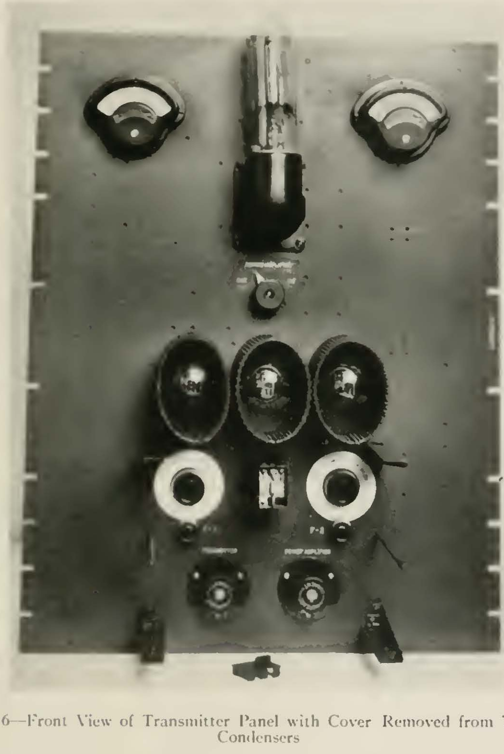


Fig. 6—Front View of Transmitter Panel with Cover Removed from Tuning Condensers

being obtained from filament drop, "B" battery drop and a combination of these two.

The transmitting unit shown in Figs. 6 and 7 is divided into two parts, the transmitting circuit proper and the power amplifier. The first is a circuit comprising a 101-D tube functioning as a Hartley oscillator with inductive feed-back, a 223-A tube operating as a speech amplifier or modulator and a 223-A tube operating as a high frequency amplifier. The plate or constant current system of modulation is employed but differs somewhat from the usual practice in

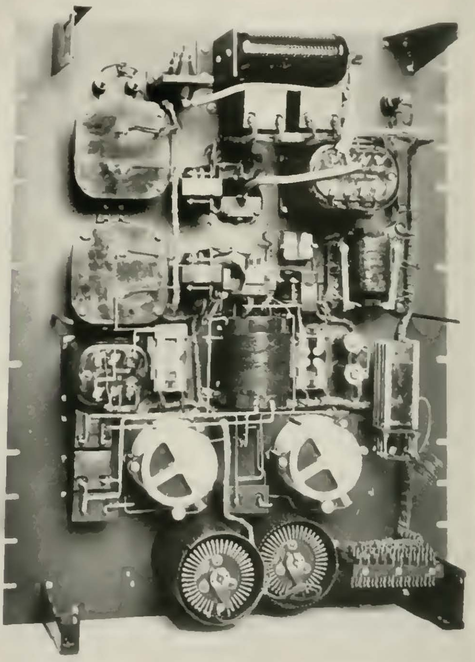


Fig. 7—Rear View of Transmitter Panel with Cover Removed

that the output of the high frequency amplifier is modulated rather than the output of the oscillator itself. This scheme was found to deliver more modulated power than the usual arrangement since it is not limited to the same extent by the overloading of the high frequency amplifier. This circuit has a power output of one watt, which has proved to be ample for normal operation of the carrier system.

To provide for operation of the system when the attenuation on the power line has been materially increased by line fault conditions a power amplifier is provided. This amplifier employs a 50 watt tube (211-A) and is placed in the circuit by a simple switching operation. When this amplifier is operated, the output of the transmitting circuit is impressed upon the grid of the 50 watt tube and amplified to approximately fifty times its normal power output.

In the present type of carrier system duplex or two way operation is secured by the use of two different carrier frequencies, one for transmission in each direction. As will be pointed out later in the

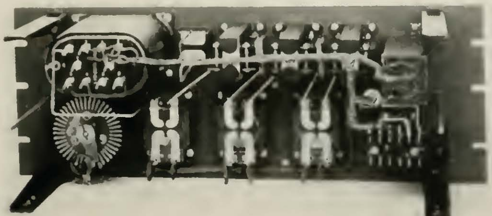


Fig. 8— Rear View of Receiver Panel

section on signaling the lower frequency is always assigned to the calling station. The transmitting circuit must therefore operate at two different frequencies. This change is accomplished by the automatic operation of the relay shown in Fig. 6. The operation of this relay changes the capacity in the oscillating circuit, thereby changing its frequency. The values of the two frequencies at which the transmitting circuit operates are determined by the variable condensers F1 and F2, Fig. 6, and certain fixed condensers which are connected in parallel with the variable condensers.

The receiving unit shown in Fig. 8 is extremely simple. It is not tuned and the only control is the filament rheostat. It consists of three 101-D vacuum tubes operating respectively as a carrier frequency amplifier, a negative grid potential detector and an audio frequency amplifier.

Two way operation is secured by operating the transmitting and receiving circuits at different frequencies and separating them by means of filters. In the single channel systems this separation is secured by a high pass filter and a low pass filter although in the multiple channel system band pass filters will be employed. Fig. 9 shows attenuation versus frequency characteristics of the high and low pass filter combination. A study of these curves shows that the transmission loss or attenuation in the high pass filter to frequencies transmitted by the low pass filter is never less than 90 T.U., which corre-

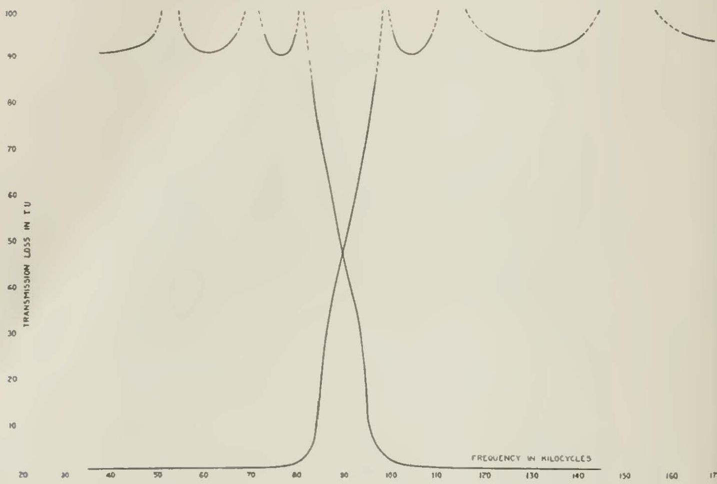


Fig. 9 Transmission Characteristic of High Pass and Low Pass Filters

ponds to a current ratio of approximately 30,000 or a power ratio of approximately 9×10^8, and the attenuation in the low pass filter to the frequencies transmitted by the high pass filter is also equal to or greater than 90 T.U.

The characteristics of these filters are remarkable when it is considered that the frequency range in which they operate is higher than that employed for multiplex carrier telephone systems, the attenuation secured is higher than that ordinarily required for such systems, and a power of 50 watts has to be transmitted through them thereby introducing special problems in the design of the coils and condensers. Figs. 10 and 11 are front and back views of one of these filters.

One of the unusual features in the use of these filters is the fact that the position of the filters in the circuit is changed from time to time by the operation of the relay shown on Fig. 11, that is to say, when the transmitting circuit is operating at a frequency lower than 80 K.C. the low pass filter is connected to it and when the transmitting circuit is operating at a frequency higher than 100 K.C. the high pass filter must be connected to it.

Signaling System

Signaling or ringing is accomplished at the transmitting end by changing the frequency of the oscillator from a frequency below 80 K.C. to a frequency above 100 K.C. without changing the filters. This is

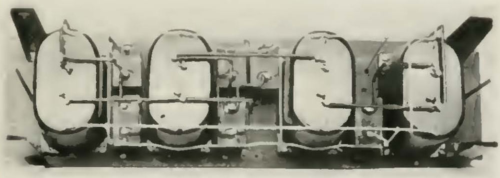


Fig. 10—Front View of Low Pass Filter with Cover Removed

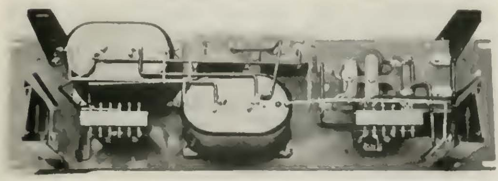


Fig. 11—Rear View of Low Pass Filter with Cover Removed

accomplished by operating and releasing the relay in the oscillator circuit. Since the filter connected to the transmitting circuit will pass only one of these frequencies, pulses of the carrier frequency are sent out on the line. At the receiving end these pulses are amplified and rectified and the change in the space current of the detector operates a marginal relay. The number and arrangement of these pulses is controlled by a spring-operated selector key of the type commonly employed for telephone dispatching on railroad lines. At the receiving end these pulses operate a train dispatching selector relay (see Fig. 12) which responds to 17 impulses. This selector relay will respond to only two arrangements of these 17 pulses. The first arrangement is 17 consecutive pulses in which case these pulses must follow one another at the correct speed and must be of the correct duration. This makes it possible to ring all stations at the same time as may be desirable in issuing general orders. The selector relay will also respond to 17 pulses broken up into three groups in which case the correct number of pulses must occur in each group and the total of the three groups must be 17. This makes it possible to

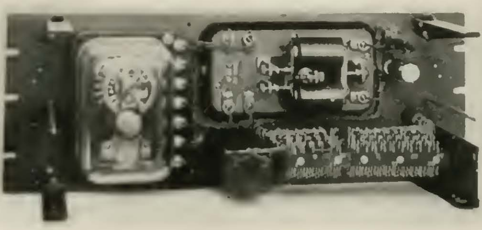


Fig. 12—Rear View of Signaling and Low Frequency Panel Showing the Signaling Apparatus

select one station from a group of more than 50 stations without disturbing the others. In addition to these desirable characteristics a single selector relay will provide selective ringing on four low frequency extensions from the carrier terminal.

The carrier equipment may be operated with complete control and talking facilities from either a telephone located at the carrier terminal or a telephone some distance from the carrier terminal but connected to it by a physical telephone circuit. In any event the control is automatic, the transmitting circuit operating only when the receiver is off the switchhook, while the receiving circuit operates continuously

Designating the carrier frequency which is below 80 K.C. as F_1 and the carrier frequency which is above 100 K.C. as F_2, the operation of a carrier system comprising three carrier terminals designated as A, B and C with a remote control station designated as A_1 located at the load dispatcher's office and separated from the carrier terminal by several miles of physical telephone circuit is as follows. Each of these stations may communicate with any of the other stations. Communication between A, B and C is carried on over carrier circuits; communication between A and A_1 is carried on over the physical circuit while communication between A_1, B and C is carried on over circuits which are composed of a carrier circuit and a physical circuit operating in tandem. When in the normal or non-operated conditions, each of these carrier terminals is set up to receive a signal on frequency F_1, but when the receiver is removed from the switch-hook at any station to initiate a call, the carrier terminal corre-

Fig. 13—110 K.V. Coupling Condensers Used for Coupling Carrier Circuit to a 110 K.V. Power Line

sponding to that telephone is automatically set up to transmit on frequency F_1 and receive on frequency F_2. When the ringing key is operated, pulses of frequency F_1 are sent out and received at all of the other carrier terminals. At the called station these pulses operate a selector relay and ring the bell, and when the operator removes his receiver from the switch-hook to answer the call, his carrier terminal is automatically set up to transmit on frequency F_2 and receive on frequency F_1. This switching of the transmitting and receiving circuits from one frequency to another is necessary where more than two stations are operated on the same system and it is desirable for every station to be able to call every other station without routing the call through a central point.

If station A_1 is connected with station A by means of two or more pairs of telephone wires which are not exposed to high voltage power

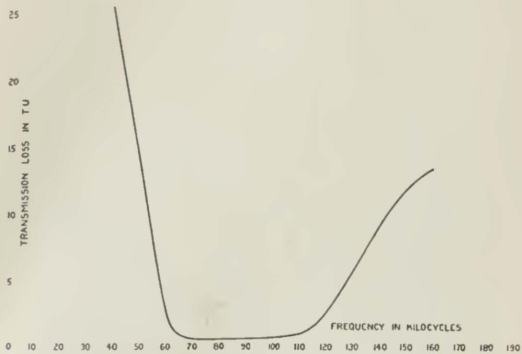


Fig. 14—Transmission Characteristic of Coupling Band Pass Filter

lines, a simple D.C. remote control circuit may be employed. However, if only two wires are available or if the telephone lines to be used are exposed to high voltage power lines and must therefore be equipped with insulating transformers and drainage coils, it is necessary to employ a somewhat more complex alternating current control circuit. In this circuit the 135 cycle interrupters and relays familiar to the telephone plant are employed.

The voice frequency circuits used in connection with this carrier equipment are the standard two wire and four wire circuits used in commercial telephone practices.

Coupling by Condensers and by Distributed Capacity

Fig. 13 shows two of the 120 K.V. coupling condensers developed by the Ohio Brass Co. Each of these condensers has a capacity of .003 μf although similar condensers having a capacity of .007 μf are also available. These condensers are approximately 5 ft. in diameter and 12 ft. high over the bushing and weigh about 8,000 pounds. The condenser element is made up of a large number of small condensers in parallel, the assembly being immersed in transformer oil.

At present these condensers are employed as the series capacity element of a single section, confluent type, Campbell band pass filter as shown by Fig. 22, the general attenuation characteristic being shown by Fig. 11. This filter is intended to transmit efficiently the carrier frequencies, and to exclude power frequency currents.

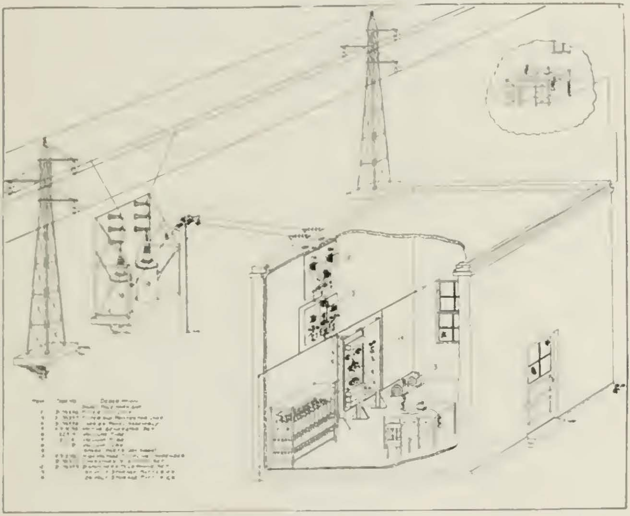


Fig. 15—Typical Layout of Power Line Carrier Telephone System, Using High Voltage Condensers for Coupling to Power Line

In Fig. 15 is shown a typical layout of a condenser coupled power line carrier telephone system.

In employing the distributed capacity type of condenser for coupling to the power line, two coupling wires (sometimes incorrectly called antennae) are suspended parallel to the power conductors for a distance of approximately 1,000 ft. Fig. 16 shows the last tower supporting the coupling wires in an installation at Anniston, Alabama. This is a twin circuit 110 K.V. power line and in order to secure coupling to both lines, the coupling wires are suspended midway between the top and bottom phases. The box shown on the tower in Fig. 16 is the coupling wire tuning unit shown in Fig. 17. The coupling wires are terminated in this tuning unit. In Fig. 18 is

Fig. 16 Distant End of Typical Coupling Wire Installation Showing Coupling Wire Tuning Unit

shown the schematic diagram of the wire coupling circuit and Fig. 19 illustrates the character of the resonant peaks secured by this circuit. The series inductances L_1 and the terminating inductance L_2 are variable and by adjusting them the points of resonance may be shifted to correct for variations in the coupling wire inductance and capacity for different installations. Fig. 20 illustrates a typical carrier terminal installation employing the wire coupling method.

The only point in favor of the wire coupling as compared with the condenser coupling is the fact that for power lines of voltages higher

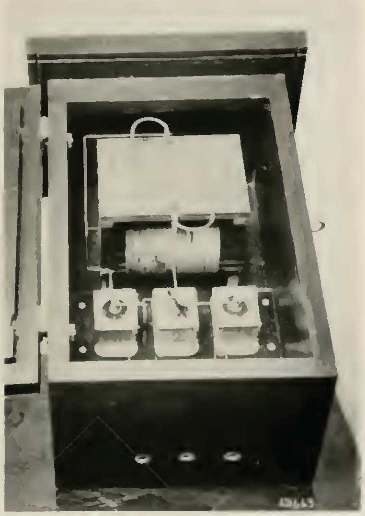


Fig. 17—Coupling Wire Tuning Unit

than 33 K.V. it is somewhat cheaper. On the other hand condenser coupling is much more efficient, thereby increasing the range and reliability of the system. It also permits high quality transmission, the transmission through it is not affected by small variations in frequency, and the component parts are of constant value determined at the time of manufacture and require no adjustment at the time of installation. In addition to these advantages the inspection and maintenance of the condenser is easier than for the coupling wires.

Protective Measures

In considering the problem of safety to the operating personnel and the equipment from the power line voltage, the normal insulation supplied by the high voltage condenser where it is employed or by the air separation where the coupling wires are employed, is disregarded, since this insulation may fail, thereby applying the power line voltage to the line terminals of the coupling circuit shown in

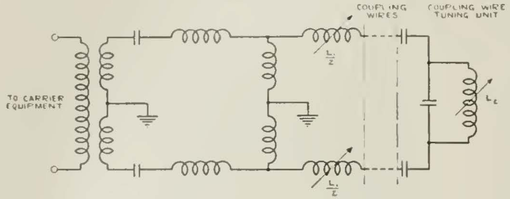


Fig. 18—Schematic of Wire Coupling Circuit

Fig. 21. The circuit shown in this figure is the same both for condenser and for wire coupling installations. The first element of protection is the horn gap, which is mounted outside of the building and serves to limit the voltage to ground which the drop wire fuse,

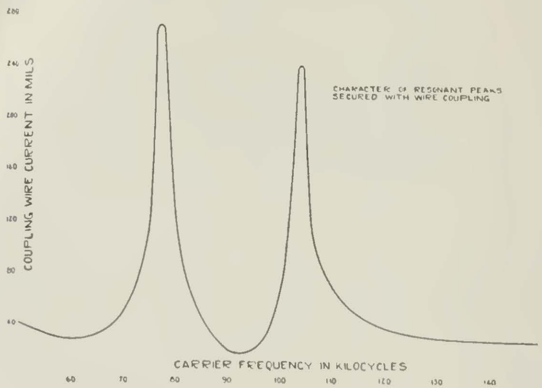


Fig. 19—Character of Resonant Peaks secured with Wire Coupling

constituting the second element of protection, will have to break. This fuse consists of an element inside of a porcelain tube the ends of which are closed by lead caps. This fuse is about 5 inches long and 1/2 inch in diameter and is supported by the wire itself. When it tails, the arc established within the porcelain tube causes the tube to break and permits the wires to fall apart. In power line carrier telephone practice this fuse is so installed that a clear drop of at least 20 ft. is obtained. The third element of protection is the shunt coil with the mid-point grounded. In many respects this element is the

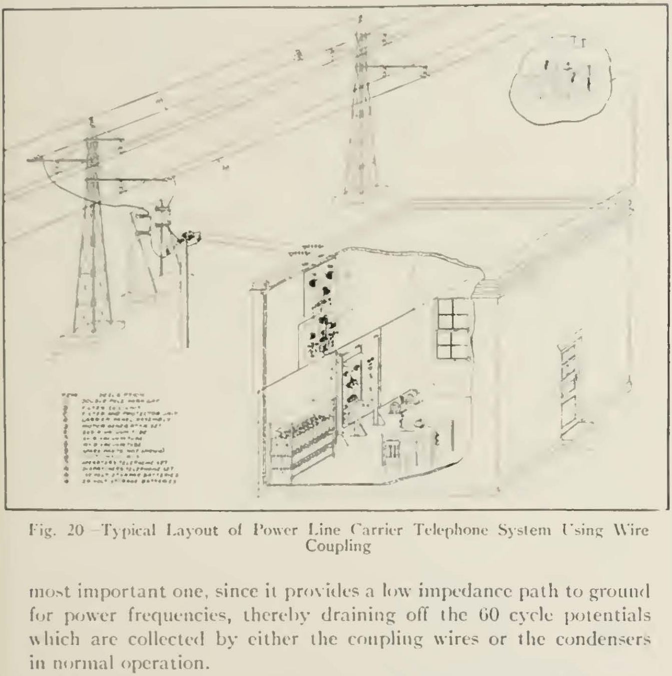


Fig. 20—Typical Layout of Power Line Carrier Telephone System Using Wire Coupling

most important one, since it provides a low impedance path to ground for power frequencies, thereby draining off the 60 cycle potentials which are collected by either the coupling wires or the condensers in normal operation.

As will be noted from Fig. 27 the line series inductances and this shunt inductance coil comprise a unit (the upper panel) which is known as the filter coil unit. The coils on this unit are insulated for 20,000 volts on the line terminals and are constructed of edgewise wound copper ribbon large enough to carry heavy momentary currents without damage. The fourth element of protection is a fused switch and surge arrester such as is commonly employed for the protection of private telephone lines exposed to power lines. This device consists of fuses in series with the line and forming the blades of a switch. These fuses have been found satisfactory for the interruption of voltages as high as 25,000. Following this fused switch is a 1,500 volt breakdown static spark gap to ground and a 500 volt breakdown vacuum gap

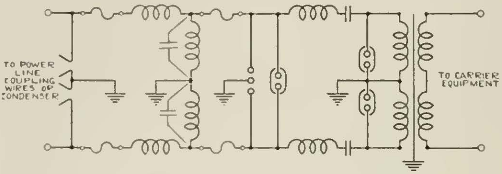


Fig. 21—Schematic of Protection Circuits

across the line. Following these there are two series capacity elements which are high voltage mica condensers. These condensers have a capacity of .007 μf. and a breakdown voltage in excess of 7,500. Finally, there is provided a repeating coil with the mid-point of the line side

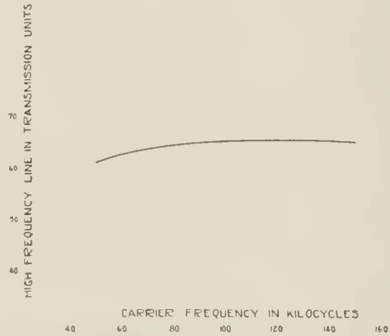


Fig. 22—Change in the Attenuation of the High Frequency Line Necessary to Maintain a Constance Voice Frequency Level with Variation in the Frequency of the Carrier

winding grounded and protected by 500 volt vacuum gaps to ground. This repeating coil is also provided with a grounded shield between the windings and has a breakdown voltage from the winding to the shield of 1,000 volts. The operation of this protective circuit has been demonstrated several times in the field by connecting one phase of a 110 K.V. power line directly to one of the line terminals of the protective circuit. In every case the circuit has operated satisfactorily. In no case has any of the standard apparatus been damaged nor has there been any evidence that the elements of protection beyond the third, that is, the shunt coil with the mid-point grounded, have been called upon to function.

Transmission Level Characteristics

Fig. 22 shows the attenuation (expressed in transmission units) of the high frequency line versus the carrier frequency of K.C. It will be noted that over the range from 50 K.C. to 150 K.C. the variation

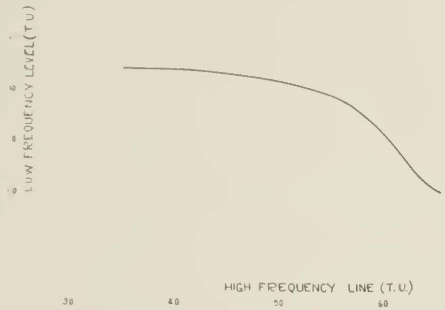


Fig. 23—Variation of Overall Gain with the Attenuation of the High Frequency Line

in attenuation is less than 5 T.U. This curve was made with a constant audio frequency input of 3.35 mils and an output of 3.35 mils from the carrier circuits, the audio frequency being 1,000 cycles. The variation of audio frequency level with the attenuation of the high frequency line is shown in Fig. 23. The observations given in Fig. 24 were made on an artificial transmission line in which the line constants, and therefore the attenuation, could be readily changed without changing the carrier frequency. The shape of this curve is a function of the receiving circuit since the audio input, carrier frequency and the modulated output of the transmitting circuit are maintained constant. It shows that for audio frequency levels lying between −10 and +10 T.U. the equivalent is approximately a straight line function of the attenuation of the high frequency line, and that therefore the receiving circuit is not overloaded.

Fig. 24 shows the audio frequency load characteristic. This curve is principally a function of the load characteristic of the modulator and it shows that for inputs greater than 1 mil, the modulator is overloaded. In practice the overloading of the modulator is prevented by increasing the average low frequency line equivalent to an attenuation of 10 T.U. by means of a resistance artificial line. This

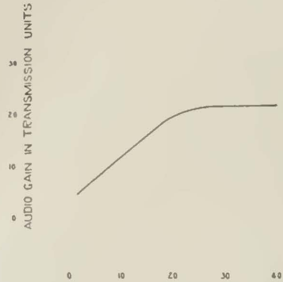


Fig. 24—Transmitting Circuit Load Characteristic

arrangement is desirable in order that the balancing of the low frequency hybrid coil may not be complicated when operating over very short physical circuits.

The curve in Fig. 25 is a single frequency quality characteristic and shows that where the method employed for connecting to the

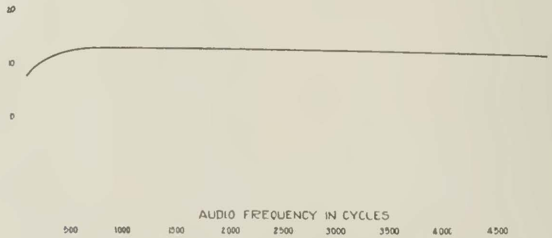


Fig. 25—Single Frequency Quality Characteristic

power line will permit, remarkably true voice transmission may be secured. The variation in the equivalent over the range from 100 cycles to 5,000 cycles is only $5\frac{1}{2}$ T.U., while the variation from 300 cycles to 5,000 cycles is only 2 T.U. Reference to Fig. 19 will indicate, however, that less satisfactory quality characteristics are obtained when the wire coupling method is employed, because of the sharpness of resonance of the coupling circuit.

Alabama Power Company Installation

Figs. 26 and 27 are photographs of the installation of power line carrier telephone equipment at the Anniston substation of the Alabama

Fig. 26—Typical Power Line Carrier Telephone Installation

Power Company. Fig. 26 illustrates the simple character of the assembled units and freedom from controls. The right hand bay is devoted to power control apparatus with space reserved for the 135 cycle remote control equipment when it is employed. The left

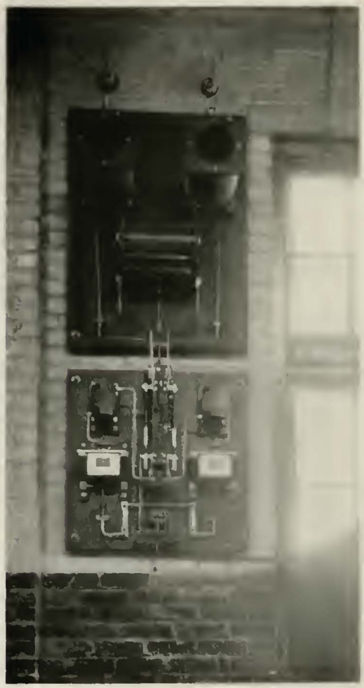


Fig. 27—Typical Installation of Coupling Panels

hand bay includes the transmitting and receiving circuits, the high and low pass carrier frequency filters and the voice frequency and D.C. control circuits. Beginning at the top of this bay, the first panel, which is blank on front, carries the system terminal. block to which all wiring except the power supply is connected. The second panel is the high pass filter; the third panel is blank. The fourth panel is the transmitting equipment, both low power and high power. The fifth panel is the receiving circuit; the sixth panel contains the voice frequency and signaling equipment. The seventh panel contains D.C. control equipment, and the bottom panel is the low pass filter. On the wall to the right of the carrier panel assembly are shown the filter coil unit and the filter and protector unit. These units are more clearly shown in Fig. 27 and diagrammatically in Fig. 21. Returning to Fig. 26, the desk stand which the operator is using is that associated with the carrier equipment, while the key mounted on the table immediately to the left of the desk stand is the selector key employed for ringing. Fig. 16 shows the coupling wire installation at this station.

The power line carrier telephone equipment which has been briefly described in the foregoing article is in successful operation today on several power systems in this country. Its reliability, simplicity of operation and maintenance have been well established.

The large number of variables which are involved in line failure conditions make it impossible to predict what effect these emergency conditions may have on the operation of the carrier equipment. The fact remains, however, that under many simulated and actual trouble conditions successful operation of the carrier equipment has been obtained.

With the growing need of power companies for communication facilities, it is probably only a question of a very short time before multiple channel carrier systems will be in operation on the large power systems of this country.

Abstracts of Bell System Technical Papers Not Appearing in the Bell System Technical Journal

Photomechanical Wave Analyzer Applied to Inharmonic Analysis.[1] C. F. Sacia. This type of Fourier Analysis deals with wave-forms which are not strictly periodic, since they are of finite duration and of varying cyclic forms. Hence in a finite frequency range they have an infinite number of infinitesimal components (shown by the Fourier Integral) as contrasted with the finite number of finite components at regular intervals (shown by the Fourier Series).

This analyzer utilizes the continual repetition of the aperiodic wave, deriving therefrom a periodic wave, the infinitesimal components neutralizing except for frequencies which are integral multiples of the frequency of repetition; here the components build up to finite magnitudes. The simple relation between these components is seen from the corresponding Fourier Integral and Series identities for the unrepeated and repeated waves respectively. By increasing the period of repetition a new set of components can be similarly derived.

The wave form is represented as a black profile on a transparent strip whose ends are joined to form an endless belt. Driven at constant speed past a transverse illuminated slit, it generates light fluctuations which are converted into electrical fluctuations by means of a selenium cell. A tuned circuit, amplifier, rectifier and microammeter are used to select and measure the components, while the frequencies are determined by the speed of the strip, the frequency of tuning, and the time scale of the original wave form.

"*Demagnetization and Hysteresis Loops.*"[2] L. W. McKeehan and P. P. Cioffi. The fact that permalloy shows its maximum initial permeability in the absence of external magnetic fields is used to check the exact compensation of the earth's magnetic field or other stray fields by measurement of the initial permeability of a strip or wire of permalloy placed parallel to the field component to be compensated. Increased accuracy is obtained by the use of somewhat greater fields than those which approximately give the initial permeability. The effect of demagnetization by an alternating current field is studied with samples of the same sort, the apparent permeability varying as the external field at the time of magnetization is varied. The dissymmetry in hysteresis loops where the upper and lower limits

[1] J. O. S., R. S. I., Vol. 9, pp. 487–491, 1924.

[2] J. O. S., R. S. I., Vol. 9, pp. 479–485, 1924.

are unsymmetrical with respect to the zero of magnetic field is illustrated and the detection of such dissymmetry is discussed.

A Classified List of Published Bibliographies in Physics, 1910–1922.[3] KARL K. DARROW. This work, undertaken at the request of the National Research Council, represents an attempt to cope with the problem of providing a convenient and adequate bibliography of physics, not by actually writing a complete classified bibliography (which would fill a huge volume and require the prolonged labor of several men), but by listing the very numerous partial bibliographies under a detailed subject-classification. Many of the accounts of research published in scientific journals contain short histories of the previous work in the subjects which they treat, many others contain lists of references, and there are also a number of critical or uncritical reviews of particular fields with thorough documentations. The *Classified List of Published Bibliographies* refers to all of these which appeared in any of the familiar physical journals between 1910 and 1922 inclusively, and a number of books as well; it is believed that almost every article upon a physical subject, which has ever been cited or reviewed in another article, can be traced through the *List.* The system of classification, in which the field of physics is divided into seventy-five classes with numerous subdivisions, is much the most detailed and elaborate which has been made out for the science of physics in a score of years. An adequate system of classification is of great value in any science, for researches which are clasified under it are not only made easy to trace, but their various aspects and their mutual relations can be emphasized. Because of the rapid growth and evolution of physics, the earlier systems have mostly become inadequate; but it is hoped to make and keep this system effective by constant attention and revision, and to extend the use of it.

Transmitting Equipment for Radio Telephone Broadcasting.[4] EDWARD L. NELSON. The general transmission considerations applying to any system for the high quality transmission of speech or music are outlined briefly, and the specific requirements to be met by the various apparatus units in a radio broadcasting equipment are discussed in some detail. The standard Western Electric 500-watt broadcasting equipment, which has found application in some fifty of the larger stations in this country and abroad, is described. Its performance capabilities are illustrated and it is indicated that a standard of performance has been attained which renders possible reproductions not substantially different from the original.

[3] Bulletin of the National Research Council, No. 47.

[4] Proc. of The Inst. of Radio Engineers, Vol. XII, page 553, 1924.

"*The Vapor Pressures of Rochelle Salt, the Hydrates of Sodium and Potassium Tartrates and Their Saturated Solutions.*"[5] H. H. Lowry and S. O. Morgan. The vapor pressures were determined by a static method at several temperatures between 15° and 40°. Temperatures were controlled to ±0.1° and the pressures read to ±0.1 mm. The measurements on the saturated solution of Rochelle salt show that the solid phase in such a solution is unstable above 40°, in agreement with other investigators.

Minimal Length Arc Characteristics.[6] H. E. Ives. This paper is a study of the electrical discharges which occur between opening contacts. It is found that the discharge occurring when currents below a certain value are broken are atmospheric sparks corresponding to a definite breakdown voltage, which in the case of air is about 300 volts. Above a critical value of current, which is different for every material, the discharge is an arc, in which the voltage corresponding to the discharge varies with current. Spectograms taken in the two regions show only the air spark spectrum for all materials below the critical current and the arc spectra of the materials above the critical current. The characteristic equations of the arcs caused by the opening contacts are derived and are used to obtain expressions for the current vs. time relations at the opening contact.

The Dependence of the Loudness of a Complex Sound Upon the Energy in the Various Frequency Regions of the Sound.[7] H. Fletcher and J. C. Steinberg. Two complex sounds were studied, one with a continuous energy frequency spectrum corresponding to connected speech, the other a test tone having discrete frequency components. By means of filters the energy was removed from all frequencies either above or below a certain frequency, and the resulting decrease in loudness was measured by attenuating the original sound without distortion until equal in loudness to the filtered sound. Taking the average results for six observers, this decrease was found to depend on the absolute values of the loudness. For a loudness of 22 units above threshold, each frequency region contributes to loudness in proportion to the energy in that region weighted according to the threshold energy for that frequency. For a loudness above 30 units, however, this is no longer true, because of the non-linear character of the response of the ear. By assuming each frequency region contributes in proportion to a fractional power of the weighted energy of that region, values of the total loudness in agreement with ob-

[5] Jour. Am. Chem. Soc., Vol. 45, pp. 2192–2196, 1924.

[6] Journal of the Franklin Institute, Vol. 198, pp. 437–474, 1924.

[7] Physical Review, Vol. 24, page 306, 1924.

served values are obtained if proper values are taken for the fractional power, decreasing to one third as the loudness increases to 100 units.

Correlation Between Crack Development in Glass While Conducting Electricity and the Chemical Composition of the Glass.[8] EARLE E. SCHUMACHER. A study was made of the susceptibility to crack development shown by five different kinds of glass when they were subjected to the action of an electric current. The results indicated that the tendency to crack increased with increasing alkali content of the glass and with increasing electrical conductivity.

Report of the Chairman of the Telegraphy and Telephony Committee of the American Institute of Electrical Engineers.[9] O. B. BLACKWELL. This report gives a brief summary of the advances which have been made or which have come into prominence in the communication art during the year. Papers which have been presented before the Institute and which, in general, have recorded such advances are reviewed.

Selective Circuits and Static Interference.[10] J. R. CARSON. This paper is an application of a general mathematical theory to the question as to the possibilities and limitations of selective circuits when employed to reduce "Static" interference. In the case of static interference and random disturbances in general the random and unpredictable character of the disturbances makes it necessary to treat the problem statistically and express the results in mean values. In spite of the meagre information available regarding the character and frequency distribution of static, this treatment of the problem yields general deductions of practical significance. The conclusion is reached that for given signal requirements there is an irreducible residue of static interference which cannot be eliminated. This limit is closely approached when a filter of only two or three sections is employed as the selective circuit, and only a negligible further gain is made possible by the most elaborate circuit arrangements. A formula is also given for calculating the relative figures of merit of selective circuits with respect to random interference.

The Guided and Radiated Energy in Wire Transmission.[11] J. R. CARSON. This is a mathematical analysis of wave propagation along guiding wires from the fundamental equations of electromagnetic theory. It is shown that the engineering theory of wire transmission is incomplete, and that, in addition to the transmitted wave of en-

[8] Jour. Am. Chem. Soc., Vol. XLVI, No. 8, August, 1924.

[9] Journal of the American Inst. of Elec. Engineers, Vol. 43, page 1083, 1924.

[10] Trans. A. I. E. E., 1924.

[11] Jour. A. I. E. E., Oct., 1924.

gineering theory, an infinite series of complementary waves exist. It is through these waves that the phenomena of radiation are directly accounted for. Except for the phenomena of radiation, however, the complementary waves are of theoretical rather than practical interest in present-day transmission practice, and except in extreme cases they may be ignored in practice without appreciable error.

Sound Magnification and Its Application to the Requirements of the Deafened.[12] HARVEY FLETCHER. A general description of the generation and propagation of sound waves was given and experiments performed to illustrate the principles involved. The general requirements for aiding persons having various amounts of deafness were outlined. The relation between the loudness of speech received by the ear in a room of average acoustic characteristics and the distance the speaker is away from the ear was illustrated by a chart. Also, a chart showing the characteristic frequency regions and loudness levels of the fundamental speech sounds, and one showing the interpretation of speech at various loudness levels by persons having various degrees of hearing, were exhibited. By means of these three charts it was shown how one could predict the amount of intelligibility which would be obtained by a person having a definitely measured amount of hearing. In particular it was pointed out that such sounds as *th*, *f*, and *v* will be the first sounds to be lost as the hearing decreases. These sounds are the easiest ones to detect by lip reading so that hearing aids and lip reading go hand in hand in aiding one who is hard of hearing to obtain the proper interpretation.

Abstract of a Telephone Transmission Reference System.[13] L. J. SIVIAN. The subject is dealt with in four parts: A—The function of a transmission reference system; B—Requirements to be met by the reference system; C—Work done on the construction and calibration of a preliminary model of the new reference system; D—Proposed future development of the new reference system in its final form to be adopted as the standard for the Bell System.

A brief discussion of the methods and apparatus entering into the general problem of rating telephone transmission is given. It is

[12] Lecture given before the Annual Conference of the American Federation of Organizations for the Hard of Hearing, Washington, D. C., Thursday, June 5, and published in Volta Review, September, 1924.

A large number of the audience who listened to this lecture were hard of hearing. A rough measurement of the amount of hearing of each of those present was made and groups arranged according to the degree of hearing. The amplification was then adjusted to each group to suit their particular needs. The results seemed to be most gratifying, as nearly everybody said that it was the first time they ever heard a public lecture of this sort without difficulty since they had become hard of hearing.

[13] Electrical Communications, Vol. III, pp. 114–126, 1924.

concluded that a physical reference system is essential, and that a mere specification of its physical operating characteristics is insufficient. The inadequacy of the reference systems now in use is pointed out.

The conditions to be aimed at in the new reference system are: I—The performance of the system and of its component parts must be specifiable in terms of quantities admitting of definite physical measurement; II—The performance of the reference system, under specified operating and atmospheric conditions, must remain constant with time; III—The reference system must be free from non-linear distortion over the range of acoustic and electric amplitudes which it must handle; IV—The frequency response over the range of speech frequencies must be as nearly uniform as possible.

Of the above, conditions I and II are regarded as the most important. It is also proposed to build auxiliary reference systems which will meet conditions I and II while falling short of III and IV. These are needed for purposes of ready comparisons with the commercial circuits commonly in use.

Contributors to this Issue

F. L. Rhodes, S.B., Massachusetts Institute of Technology, 1892; American Bell Telephone Company; Outside Plant Engineer, American Telephone and Telegraph Company, 1909–19; Outside Plant Development Engineer, 1919—. Mr. Rhodes has had an active part in the development and standardization of materials, apparatus and practices employed in the underground and overhead wire plant of the Bell System. He has written many articles, among which may be mentioned those on "The Telephone" in the Encyclopedia Americana and Nelson's Encyclopedia.

George Crisson, M.E., Stevens Institute of Technology, 1906; instructor in Electrical Engineering, 1906–10. American Telephone and Telegraph Company, Engineering Department, outside plant division, 1910–14; transmission and protection division, 1914–19; Development and Research Department, transmission development division, 1919—.

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.

K. S. Johnson, A.B., Harvard University, 1907; Graduate School of Applied Arts and Sciences, 1907–09; Engineering Department of the American Telephone and Telegraph Company, 1909–13; Engineering Department, Western Electric Co., Inc., 1913–24; Bell Telephone Laboratories, Inc., 1925—. Mr. Johnson's work has related especially to the theoretical aspects of telephone and telegraph transmission.

Timothy E. Shea, S.M., Massachusetts Institute of Technology, 1919; instructor in Electrical Engineering and Physics, 1918–20; Manufacturing Department, Western Electric Company, 1920–21; Engineering Department, 1921–24; Apparatus Development Department, Bell Telephone Laboratories, 1925—. Mr. Shea has been principally engaged in the development of electric wave filters and allied apparatus.

Karl K. Darrow, S.B., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., in physics and mathematics, University of Chicago, 1917; Engineering Department, Western Electric Company, 1917–24, Bell Telephone Laboratories, Inc., 1925– . Mr. Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

W. V. Wolfe, B.S., Carnegie Institute of Technology, 1918; Signal Corps, 1918–19; General Electric Company, 1919; Standard Underground Cable Company, 1920; Engineering Department, Western Electric Company, 1920–24; Bell Telephone Laboratories, Inc., 1925– . Mr. Wolfe has been engaged in the development of various types of carrier systems.

The Bell System Technical Journal

April, 1925

The Transmission of Pictures Over Telephone Lines

By H. E. IVES and J. W. HORTON, *Bell Tel. Lab. Inc.*
R. D. PARKER and A. B. CLARK, *Amer. Tel. & Tel. Co.*

Introduction

THE problem of directly transmitting drawings, figures and photographs from one point to another by means of electricity has long attracted the attention and curiosity of scientists and engineers.[1] The broad principles of picture transmission have been recognized for many years. Their reduction to successful practice, however, required, among other things, the perfection of methods for the faithful transmission of electrical signals to long distances, and the development of special apparatus and methods which have become a part of the communication art only within the last few years. Prominent among the newer developments which have facilitated picture transmission are the photoelectric cell, the vacuum tube amplifier, electrical filters, and the use of carrier currents.

None of the systems heretofore devised have been sufficiently developed to meet the requirements of modern commercial service. The picture transmission system described in this article has been designed for practical use over long distances, employing facilities of the kind made available by the network of the Bell System.

The desirability of adding picture transmission facilities to the other communication facilities offered by the Bell System seems now to be well assured. Various engineers of the System have made suggestions and carried out fundamental studies of the possibilities for picture transmission offered by the telephone and telegraph facilities in the Bell System Plant which have aided materially in the development of the method to be described.

[1] A comprehensive account of earlier work in Picture Transmission will be found in "Telegraphic Transmission of Pictures," T. Thorne Baker, Van Nostrand, 1910, and the "Handbuch der Phototelegraphie und Telautographie," Korn and Glatzel, Leipzig, Nemnich, 1911.

The account of the picture transmission system which follows is intended to give only a general idea of the work as a whole. A number of engineers have collaborated in this work, and it is expected that later publications will describe various features of the system and its operation in greater detail.

General Scheme of Picture Transmission

Reduced to its simplest terms, the problem of transmitting a picture electrically from one point to another calls for three essential elements: The first is some means for translating the lights and shades of the picture into some characteristic of an electric current;

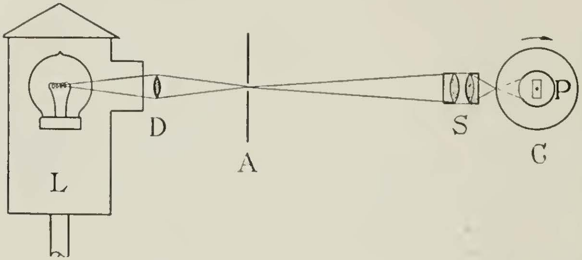


Fig. 1—Sending end optical system in section: (L) light source; (D) condensing lens; (A) diaphragm; (S) projection lens; (C) transparent picture film in cylindrical form; (P) photoelectric cell

the second is an electrical transmission channel capable of transmitting the characteristic of the electric current faithfully to the required distance; the third is a means for retranslating the electrical signal as received into lights and shades, corresponding in relative values and positions with those of the original picture.

Analyzed for purposes of electrical transmission, a picture consists of a large number of small elements, each of substantially uniform brightness. The transmission of an entire picture necessitates some method of traversing or scanning these elements. The method used in the present apparatus is to prepare the picture as a film transparency which is bent into the form of a cylinder. The cylinder is then mounted on a carriage, which is moved along its axis by means of a screw, at the same time that the film cylinder is rotated. A small spot of light thrown upon the film is thus caused to traverse the entire film area in a long spiral. The light passing into the interior of the cylinder then varies in intensity with the transmission or tone value of the picture. The optical arrangement by which a small spot of light is projected upon the photographic transparency is shown in section in Fig. 1.

The task of transforming this light of varying intensity into a variable electric current is performed by means of an alkali metal

Fig. 2—Photograph of photoelectric cell of type used in picture transmission

photoelectric cell. This device, which is based on the fundamental discovery of the photoelectric effect by Hertz, was developed to a high degree of perfection by Elster and Geitel. It consists of a vacuum tube in which the cathode is an alkali metal, such as potassium. Under illumination, the alkali metal gives off electrons, so that when the two electrodes are connected through an external circuit, a current flows. This current is directly proportional to the intensity of the illumination, and the response to variations of illumination is practically instantaneous. A photograph of a photoelectric cell of the type used in the picture transmission apparatus is shown in Fig. 2. This cell is placed inside the cylinder formed by the photographic transparency which is to be transmitted, as shown in Fig. 1. As the film cylinder is rotated and advanced, the illumination of the cell and consequently the current from it registers in succession the brightness of each elementary area of the picture.

Assuming for the moment that the photoelectric current, which is a direct current of varying intensity, is of adequate strength for successful transmission, and that the transmission line is suitable for

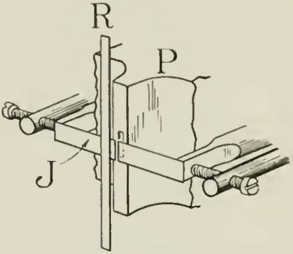


Fig. 3—Light valve details: (R) ribbon carrying picture current; (P) pole piece of magnet; (J) jaws of aperture behind ribbon

carrying direct current, we may imagine the current from the photoelectric cell to traverse a communication line to some distant point. At the distant point it is necessary to have the third element above mentioned, a device for retranslating the electric current into light and shade. This is accomplished in the present system by a device, due in its general form to Mr. E. C. Wente, termed a "light valve." This consists essentially of a narrow ribbon-like conductor lying in a magnetic field in such a position as to entirely cover a small aperture. The incoming current passes through this ribbon, which is in consequence deflected to one side by the inter-action of the current with the magnetic field, thus exposing the aperture beneath. Light passing through this aperture is thus varied in intensity. If it then falls upon a photographic sensitive film bent into cylindrical form, and rotating in exact synchronism with the film at the sending end, the film will be exposed by amounts varying in proportion to the lights and shades of the original picture. The ribbon and aperture of the light valve are shown diagrammatically in Fig. 3. Fig. 1

shows a section of the receiving end of a system of the sort postulated, with its light source, the light valve, and the receiving cylinder.

Adaptation of Scheme to Telephone Line Transmission

The simple scheme of picture transmission just outlined must be modified in order to adapt it for use on commercial electrical communication systems, which have been developed primarily for other purposes than picture transmission. Of existing electrical means of communication, which include land wire systems (telegraph and telephone), submarine cable, and radio, the wire system, as developed

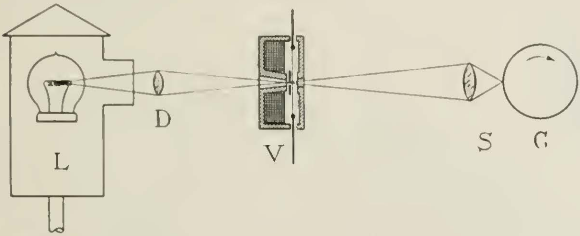


Fig. 1—Section of receiving end optical system: (L) light source; (D) condensing lens; (V) light valve; (S) projection lens; (C) sensitive film

for the telephone, offers great advantage when all factors are considered, including constancy, freedom from interference and speed. The picture transmission system has accordingly been adapted to it.

In the simple scheme of picture transmission outlined in the preceding section, the photoelectric cell gives rise to a direct current of varying amplitude. The range of frequency components in this current runs from zero up to a few hundred cycles. Commercial long distance telephone circuits are not ordinarily arranged to transmit direct or very low frequency currents, so the photoelectric currents are not directly transmitted. Moreover, these currents are very weak in comparison with ordinary telephone currents. On account of these facts, the current from the photoelectric cell is first amplified by means of vacuum tube amplifiers[2] and then is impressed upon a vacuum tube modulator jointly with a carrier current whose frequency is about 1,300 cycles per second. What is transmitted over

[2] For a very full description of the standard telephone repeater the reader is referred to "Telephone Repeaters," Gherardi and Jewett, Trans. A. I. E. E., Nov., 1919, Vol. 38, part 2, pp. 1287–1345.

Fig. 5—Portion of transmitted picture of variable width line type, enlarged

the telephone line is, then, the carrier wave [3] modulated by the photoelectric wave so that the currents, in frequency range and in amplitude, are similar to the currents corresponding to ordinary speech.

When the carrier current, modulated according to the lights and shades of the picture at the sending end, traverses the ribbon of the light valve at the receiving end, the aperture is opened and closed with each pulse of alternating current. The envelope of these pulses follows the light and shade of the picture, but the actual course of

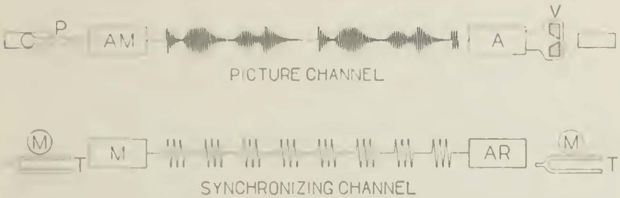


Fig. 6—Diagrammatic representation of the picture and synchronizing currents. (P) photoelectric cell; (AM) amplifier modulator; (A) amplifier; (V) light valve; (M) phonic wheel motors; (T) tuning forks; (AR) amplifier rectifier

the illumination with time shows a fine structure, of the periodicity of the carrier. This is shown by the enlarged section of a picture, Fig. 5; in this the black lines are traces of the image of the light valve aperture. Superposed on the larger variations of width, which are proportional to the light and shade of the picture, small steps will be noted (particularly where the line width varies rapidly); these are caused by the carrier pulses.

Synchronization

In order that the light and shade traced out on the receiving cylinder shall produce an accurate copy of the original picture, it is necessary that the two cylinders rotate at the same uniform rate. This, in general, demands the use of accurate timing devices. The means employed in the present apparatus consist of phonic wheels or impulse motors controlled by electrically operated tuning forks.[4] Were it

[3] A description of electrical communication by means of carrier currents will be found in "Carrier Current Telephony and Telegraphy," Colpitts and Blackwell, Trans. A. I. E. E., 1921, Vol. 40, pp. 205–300. A discussion of the relations between the several components of the signal wave employed in carrier is given in "Carrier and Sidebands in Radio Transmission," Hartley, Proc. I. R. E., Feb., 1923, Vol. 11, No. 1, pp. 34–55.

[4] A detailed description of the construction and operation of the impulse motor and its driving fork is given in "Printing Telegraph Systems," Bell Trans. A. I. E. E., 1920, Vol. 39, Part 1, pp. 167–230.

possible to have two forks at widely separated points running at exactly the same speed, the problem of synchronizing would be immediately solved. Actually this is not practical, since variations of speed with temperature and other causes prevent the two forks from operating closely enough together for this purpose. If the two cylinders are operated on separate forks, even though each end of the apparatus runs at a uniform rate, the received picture will, in general, be skewed with respect to the original. The method by which this difficulty has been overcome in the present instance is due to Mr. M. B. Long. Fundamentally the problem is solved by controlling the phonic wheel motors at each end by the same fork. For this purpose it has been found desirable to transmit to the receiving station impulses controlled by the fork at the sending end. The problem of transmitting both the fork impulses and the picture current simultaneously could be solved by the use of two separate circuits. If this were done the currents going over the two lines would be substantially as shown in Fig. 6, where the upper curve represents the modulated picture carrier for two successive revolutions of the picture cylinder, and the lower curve shows the synchronizing carrier current modulated by the fork impulses.

It would not, however, be economical to use two separate circuits for the picture and synchronizing channels, consequently the two currents are sent on the same circuit. In order to accomplish this, the picture is sent on the higher frequency carrier, approximately 1,300 cycles per second, and the synchronizing pulses are sent on the lower frequency carrier, approximately 400 cycles per second, both lying in the range of frequencies readily transmitted by any telephone circuit. These carrier frequencies are obtained from two vacuum tube oscillators.[5] The two currents are kept separate from each other by a system of electrical filters at the sending and receiving ends, so that while the current on the line consists of a mixture of two modulated frequencies, the appropriate parts of the receiving apparatus receive only one carrier frequency each.[6]

[5] The vacuum tube oscillator as a source of carrier current is described in Colpitts and Blackwell, *Loc. Cit.* A general discussion of the vacuum tube oscillator is given in the "Audion Oscillator," Heising, *Jour. A. I. E. E.* April and May, 1920. A discussion of the arrangement of the particular oscillator used with the picture transmission equipment is given in "Vacuum Tube Oscillator," Horton, *Bell System Tech. Jour.* July, 1924, Vol. 3, No. 3, pp. 508–524.

[6] The application of wave filters to multi-channel communication systems is discussed in Colpitts and Blackwell, *Loc. Cit.* More complete discussions are to be found in: "Physical Theory of Electric Wave Filters," Campbell, *Bell System Tech. Jour.* Nov., 1922, Vol. 1, No. 2, pp. 1–32.

Description of Apparatus

Mechanical Arrangements

The essential parts of the mechanism used for rotating and advancing the cylinder at the sending station, and for holding the photoelectric cell and the amplifying and modulating system are shown in the photograph, Fig. 7. At the extreme left is the phonic wheel impulse motor, which drives the lead screw through a spiral gear.

Fig. 7—Sending end apparatus showing motor, film carriage, optical system and amplifier modulator

The spiral gear ordinarily turns free of the lead screw, but may be engaged with it by a spring clutch. The lamp housing, which provides the illumination for the photoelectric cell, is in the foreground at the center of the photograph. The photoelectric cell is in a cylindrical case at the left end of the large box shown on the track and projects into the picture cylinder on which a film is in process of being clamped. The amplifier and modulator system is carried in the large box to the right, which is mounted on cushion supports to eliminate disturbances due to vibration.

The receiving end mechanism for turning and advancing the cylinder is similar to that at the sending end. The parts peculiar to the receiving end are shown in Fig. 8. They consist of the light valve, which is in the middle of the photograph, and the lens for projecting the light from it upon the cylinder. The metal cylinder

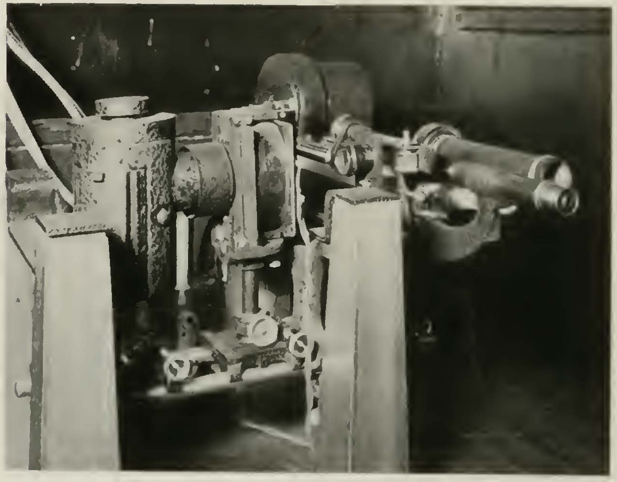


Fig. 8—View of receiving end apparatus showing light valve and observation microscope

around which the sensitive photographic film is wrapped, appears at the extreme right. The microscope and prism shown are used for inspecting the light valve aperture for adjusting purposes.

Electrical Circuits

The essential parts of the electrical circuits used are shown in the schematic diagrams, Figs. 9 and 10, in which the various elements which have been described previously are shown in their relations to each other.

Certain portions of the electrical circuits deserve somewhat detailed treatment. One of these is the amplifier-modulator system for the picture channel, the other is the filter system employed for separating the picture and synchronizing channels.

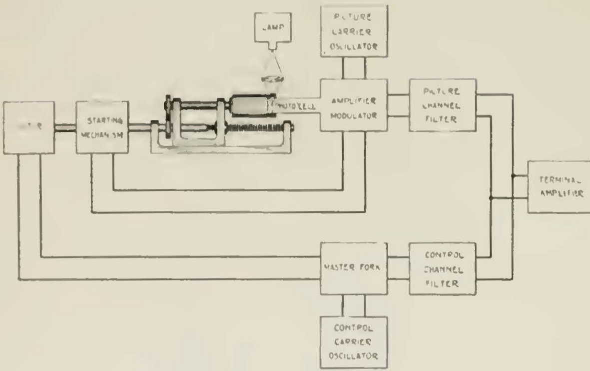


Fig. 9—Schematic diagram of sending end apparatus

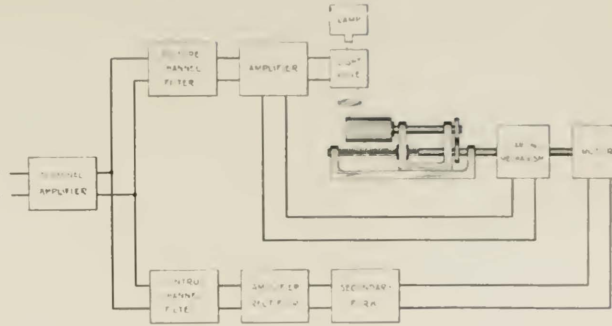


Fig. 10—Schematic diagram of receiving end apparatus

In Fig. 11 is shown (at the top) a diagram of the direct current amplifier and the modulator used for the picture channel, together with diagrams (at the bottom) showing the electrical characteristics of each element of the system. Starting at the extreme left is the

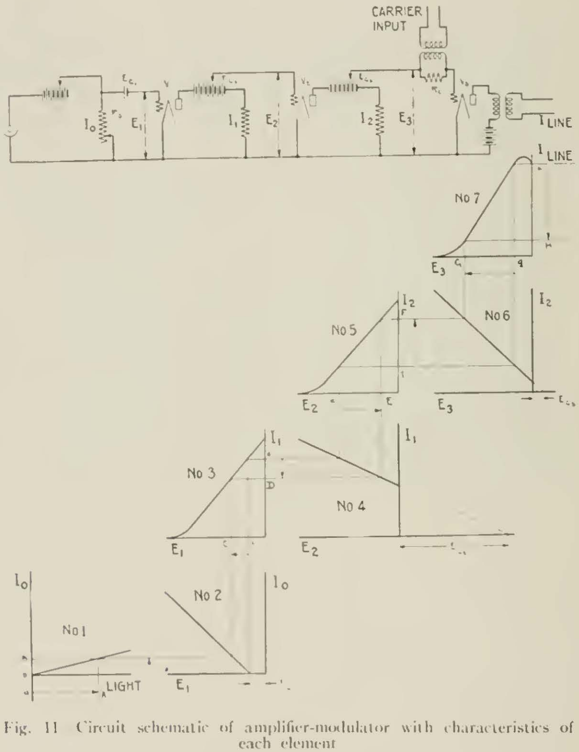


Fig. 11—Circuit schematic of amplifier-modulator with characteristics of each element

photoelectric cell, the current from which passes through a high resistance. The potential tapped off this resistance (of the order of 30 or 40 millivolts) is applied to the grid of the first vacuum tube amplifier. The second vacuum tube amplifier is similarly coupled with the first, and the vacuum tube modulator in turn to it. The relationship between illumination and current in the photoelectric cell is, as shown in diagram No. 1, linear from the lowest to the highest values of illumination. The voltage-current (*E* versus *I*) character-

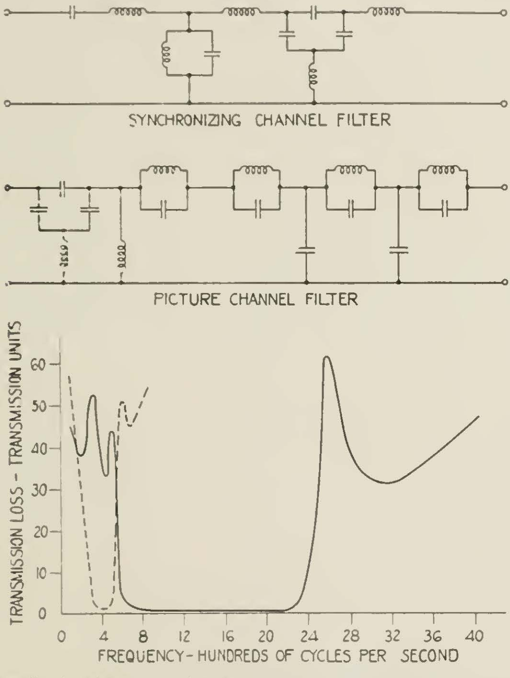


Fig. 12—Circuit schematics (above) and attenuation characteristics (below) of picture (full line) and synchronizing (dashed line) channel filters

istics of the amplifying tubes and the modulating tube circuits are shown in the figure by the diagrams which lie immediately below these tubes. They are not linear over their whole extent. It becomes necessary, therefore, in order to preserve the linear characteristic, which is essential for faithful picture transmission, to locate the range of variation of current in each of the latter tubes on a linear portion of their characteristics. This is accomplished by appropriate biasing voltages (Eg), as shown. As a consequence of this method of utilizing the straight line portions of the tube characteristics, the current received at the far end of the line does not vary between zero and finite value, but between two finite values. This electrical bias is exactly matched in the light valve by a mechanical bias of the jaws of the valve opening.

Fig. 12 shows diagrammatically the form of the band pass filters used for separating the picture and synchronizing channels, together with the transmission characteristics of the filters. The synchronizing channel filter transmits a narrow band in the neighborhood of 400 c. p. s., the picture channel filter a band between 600 and 2,500 c. p. s.

In addition to the main circuits which have been discussed, arrangements are made for starting the two ends simultaneously and for the transmission of signals. These functions are performed by the interruption of the picture current working through appropriate detectors and relays. Testing circuits are also provided for adjusting the various elements without the use of the actual transmission line.

The Transmission Line

In view of the fact already emphasized, that the currents used in picture transmission are caused to be similar both as to frequency and amplitude to those used in speech transmission, it follows that no important changes in the transmission characteristics of the telephone line are called for. With regard to the frequency range of the alternating currents which must be transmitted and also the permissible line attenuation, the transmission of pictures is less exacting on the telephone line than is speech transmission. In certain other respects, however, the requirements for picture transmission are more severe. For speech, the fundamental requirement is the intelligibility of the result, which may be preserved even though the transmission varies somewhat during a conversation. In the case of picture transmission, variations in the transmission loss of the line, or noise appearing even for a brief instant during the several minutes required for transmission are all recorded and presented to view as blemishes in the finished picture. Picture transmission circuits must, therefore, be carefully designed and operated so as to reduce the possibility of such disturbances. In transmitting pictures over telephone lines, it is also necessary to guard against certain other effects, including transient

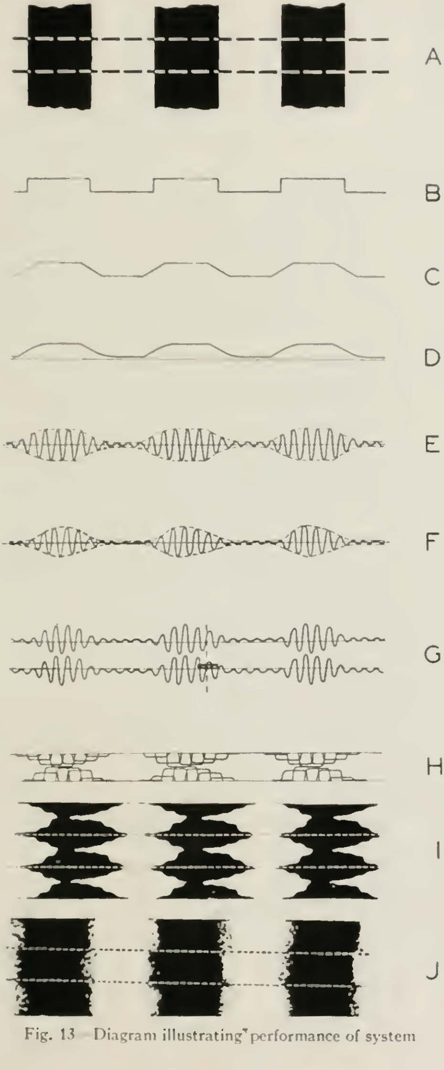


Fig. 13 Diagram illustrating performance of system

effects and "echoes" caused by reflections from impedance irregularities. A high degree of balance between the lines and their balancing networks at repeater points is also required. These conditions can be satisfactorily met on wire telephone lines. Radio communication channels are inherently less stable and less free from interference, and special means to overcome their defects are required in order to secure high-grade pictures.

Characteristics of Received Pictures

All electrically transmitted pictures have, as a result of the processes of scanning at the sending and receiving ends, a certain amount of structure, on the fineness and character of which depends the detail rendering of the result.

The origin and nature of the microscopic structure characteristic of pictures transmitted by the present process is illustrated by the diagrammatic presentation of Fig. 13, which may serve at the same time to give a review of the whole process. We will assume that the original picture consists of a test object of alternating opaque and transparent lines. Such a set of lines is shown at *A*. The lines are assumed to be moving from left to right across the spot of light falling on the film. The width of the spot of light (corresponding to the pitch of the screw) is represented by the pair of dashed lines. If the spot of light were infinitely narrow in the direction of motion of the picture film, the photoelectric current would be represented in magnitude in the manner shown at *B*. Actually the spot must have a finite length, so that the transitions between the maximum and minimum values of current are represented by diagonal lines as shown at *C*. Due to the unavoidable reactances in the amplifying system, there is introduced a certain rounding off of the signal so that the variation of potential impressed on the modulator tube follows somewhat the course shown at *D*. The alternating current introduced by the vacuum tube oscillator is, then, given the characteristics shown at *E*, the envelope being a close copy of *D*. Passing out to the transmission line, the fact that the band of frequencies transmitted by a telephone line is limited in extent results in a certain further rounding off of the envelope of the picture current as shown in *F*. The ribbon of the light valve when traversed by the alternating current from the line performs oscillations to either side of the center of the aperture, consequently opening first one side of the aperture and then the other. The two curves of sketch *G* represent the excursions of the light valve ribbon, with time, past the

Fig. 14—Example of electrically transmitted news picture—variable width line system—President and Mrs. Coolidge

edges of the aperture, which latter are indicated by parallel straight lines. Owing to the fact that the light valve aperture must have a finite length in the direction of rotation of the cylinder (indicated by the small rectangle in the center of the sketch), there is a certain overlapping of the light pulses on the film. (This is, in fact, necessary for the production of solid blacks.) These are indicated diagrammatically at *H*. In sketch I are shown, from an actual photomicrograph, the variations in the image of the light valve as traced out on the moving photographic film. Here the dashed lines represent the limits of the image as formed by one rotation of the receiving cylinder. It will be noted that the images due to the opening of the light valve in each direction form a double beaded line. These double lines are juxtaposed, so that the right hand image due to one rotation of the cylinder backs up against the left hand image due to the next rotation, thus forming on the film a series of approximately symmetrical lines of variable width. These are exhibited clearly in the enlarged section of a picture, Fig. 5. It will be understood that for purposes of illustration, the grating used as the test object in the preceding discussion has been represented as traversing the spot of light at the sending end at such a high speed that the final picture is close to the limit of the resolving power of the system. Thus the photomicrograph shown in I must be viewed from a considerable distance in order that its difference in structure from the original object *A* will disappear. A practical problem in the design of picture transmission apparatus is to so choose the speed of rotation of the cylinder with reference to the losses in resolving power incident to transmission that definition is substantially the same along and across the constituent picture lines.

There are, in general, two methods by which a transmitted picture may be received. One of these is to form an image of the light valve aperture on the sensitive photographic surface. When this is done, in the manner described in connection with Fig. 13 the picture is made up of lines of constant density and varying width. A picture of this sort is shown in Fig. 14. A merit of this kind of picture (when received in negative form) is that if the structure is of suitable size (60 to 65 lines to the inch) it may be used to print directly on zinc and thus make a typographic printing plate similar to the earlier forms of half tone, whereby the loss of time usually incident to copying a picture for reproduction purposes may be avoided. A disadvantage of this form of picture is that it does not lend itself readily to retouching or to change of size in reproduction.

Another method of picture reception is to let the light from the

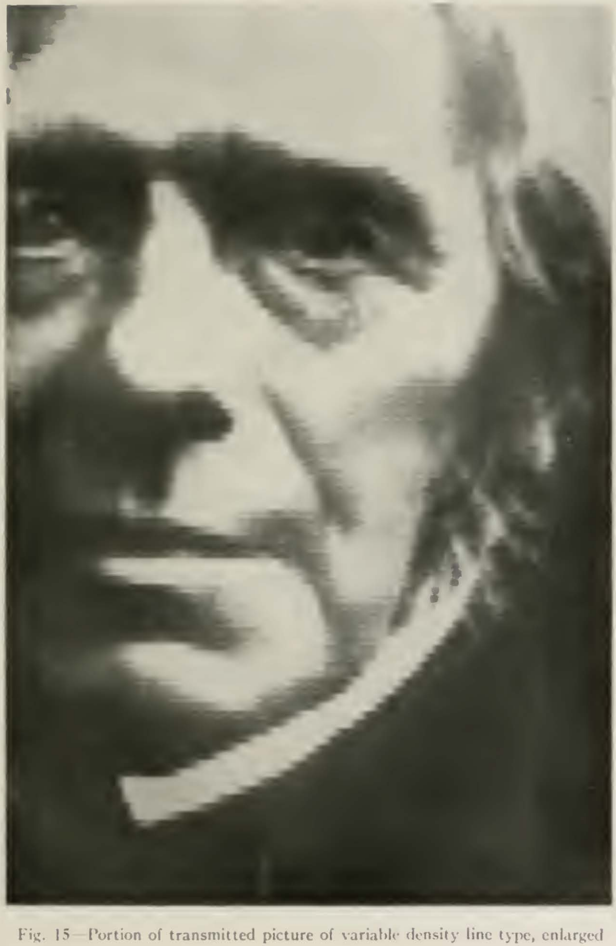


Fig. 15—Portion of transmitted picture of variable density line type, enlarged

Fig. 16—Variable density line picture—Cleveland high level bridge

Fig. 17—Variable density line picture—Portrait of Michael Faraday

Fig. 18—Variable density line picture—President Coolidge taking the oath of office, March 4, 1925

light valve fall upon the film in a diffused manner through an aperture of fixed length so that lines of constant width (exactly juxtaposed) but of varying density are produced. A photomicrograph of a variable density picture of the opaque line test object previously discussed is shown at *J*, Fig. 13. Prints made from film negatives received in this way, if the structure is chosen fine enough (100 lines to the inch or more) are closely similar in appearance to original photographic prints and may be reproduced through the ordinary half-tone cross-line screen. They may be retouched or subjected to special photographic procedures in any way desired. An enlargement of a portion of a variable density picture is shown in Fig. 15 and examples of complete pictures so received are shown in Figs. 16, 17 and 18.

Electrically transmitted pictures are, in general, suitable for all purposes for which direct photographic prints are used. Such uses include half-tone reproduction for magazines and newspapers, lantern slides, display photographs, etc. Among these uses may be mentioned, as of some interest, the transmission of the three black and white records used for making three-color printing plates. The frontispiece to this article is an example of a three-color photograph transmitted in the form of three black and white records, each corresponding to one of the primary colors, from which printing plates were made at the receiving end.

Some practical details of the procedure followed in the transmission of pictures by the apparatus described may serve to clarify the foregoing description. The picture to be transmitted is usually provided in the form of a negative, which is apt to be on glass and of any one of a number of sizes. From this a positive is made on a celluloid film of dimensions 5″ x 7″, which is then placed in the cylindrical film-holding frame at the sending end. Simultaneously an unexposed film is placed on the receiving end. Adjustments of current values for "light" and "dark" conditions are then made, over the line; after which the two cylinders are simultaneously started by a signal from one end. The time of transmission of a 5″ x 7″ picture is, for a 100 line to the inch picture, about seven minutes. This time is a relatively small part of the total time required from the taking of the picture until it is delivered in the form of a print. Most of this total time is used in the purely photographic operations. When these are reduced to a minimum by using the negative and the sending end positive while still wet, and making the prints in a projection camera without waiting for the received negative to dry, the overall time is of the order of three-quarters of an hour.

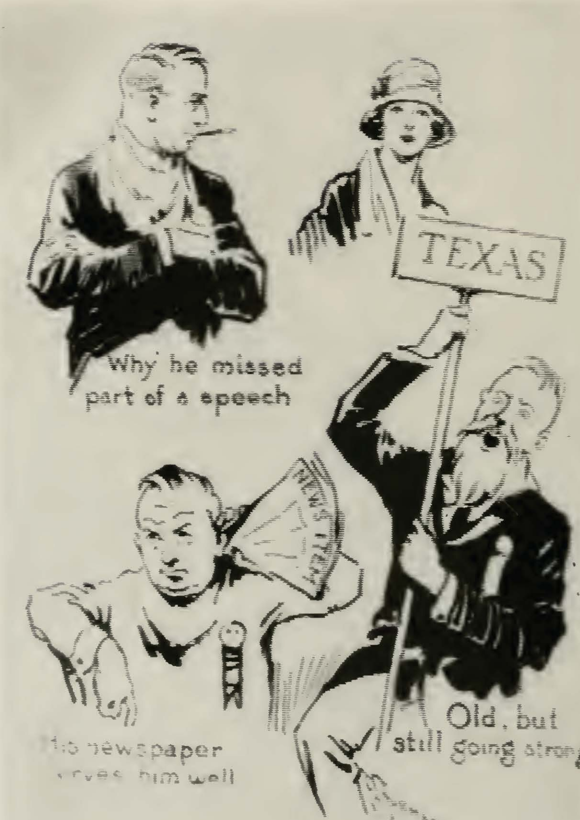


Fig. 19—Electrical transmission of cartoon

Fields of Usefulness

The fields in which electrically transmitted pictures may be of greatest service are those in which it is desired to transmit information which can only be conveyed effectively, or at all, by an appeal to vision. Illustrations of cases where an adequate verbal description is almost impossible, are portraits, as, for instance, of criminals

Fig. 20—Electrically transmitted fingerprint

or missing individuals; drawings, such as details of mechanical parts, weather maps, military maps, or other representations of transient conditions.

The value of electrically transmitted pictures in connection with police work has been recognized from the earliest days of experiments in the transmission of pictures. Besides the transmission of portraits of wanted individuals to distant points, there is now possible the transmission of finger prints. Some of the possibilities of the latter were demonstrated over the New York-Chicago picture sending circuit at the time of the Democratic Convention, July, 1924. The Police Department of New York selected the fingerprint of a criminal whose complete identification data were on file in the Police Department in Chicago. This single fingerprint, together with a code description of the prints of all the fingers, was transmitted to Chicago and identified by the Chicago experts almost instantly. This method of identification will be, it is thought, of value in those cases where difficulty is now experienced in holding a suspect long enough for identification to be completed. Fig. 20 shows a transmitted fingerprint.

The fact that an electrically transmitted picture is a faithful copy of the original, offers a field of usefulness in connection with the

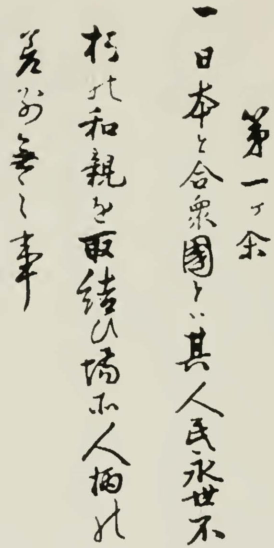

第一ヶ条
一 日本と合衆国とハ其人民永世不
朽の和親を取結ひ場所人柄の
差別無之事

Fig. 21—Transmission of autograph material—First section of Japanese-American Treaty of 1853

transmission of original messages or documents in which the exact form is of significance, such as autographed letters, legal papers, signatures, etc. It would appear that this method might under certain circumstances save many days of valuable legal time and the accumulation of interest on money held in abeyance. For these reasons, it is thought that bankers, accountants, lawyers, and large real estate dealers will find a service of this kind useful. Fig. 22 illustrates the transmission of handwriting.

Messages in foreign languages, employing alphabets of forms not suited for telegraphic coding, are handled to advantage. Thus, Fig. 21 shows the first section of the original Japanese-American treaty in Japanese script, as transmitted from New York to Chicago.

Advertising material, particularly when in the form of special typography and drawings is often difficult and costly to get to distant publishers in time for certain issues of periodicals and magazines. A wire service promises to be of considerable value for this purpose.

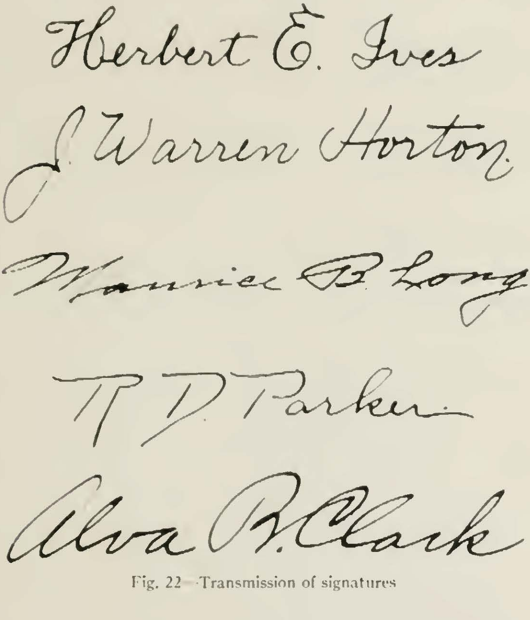


Fig. 22—Transmission of signatures

A very large field for electrically transmitted pictures is, of course, The Press. Their interest in the speedy transportation of pictures has been indicated in the past by the employment of special trains, aeroplanes, and other means for quickly conveying portraits and pictures of special events, to the large news distributing centers. The use of pictures by newspapers seems at present to be growing in favor, and many are now running daily picture pages as regular features.

Some of the possibilities in this direction were demonstrated by the picture news service furnished to newspapers, especially those in New York and Chicago, during the 1924 Republican and Democratic National Conventions at Cleveland and New York. During these conventions several hundred photographs were transmitted between Cleveland and New York and between New York and Chicago, and copies furnished the Press at the receiving points. Photographs made shortly after the opening sessions, usually about noon, were transmitted to New York and Chicago and reproduced in afternoon papers. A demonstration of picture news service on a still larger scale was furnished on March 4th, 1925, when pictures of the inauguration of President Coolidge were transmitted from Washington simultaneously to New York, Chicago and San Francisco, appearing in the afternoon papers in all three cities. Illustrations of typical news pictures are given in Figs. 14 and 18. The transmission of timely cartoons offers another field for service, Fig. 19.

Other news-distributing agencies can also use electrically transmitted pictures to advantage. Among these are the services which make a specialty of displaying large photographs or half-tone reproductions in store windows and other prominent places. Electrically transmitted pictures of interesting events, about which newspapers have published stories, appear suited to this service, and have already been so used by some of these picture service companies. They may also be used as lantern slides for the display of news events of the day by projection either upon screens in front of newspaper offices or in moving picture theaters.

Miscellaneous commercial uses have been suggested. Photographs of samples or merchandise, of building sites, and of buildings for sale may be mentioned. The quick distribution of moving picture "stills" which is now done by aeroplane is one illustration of what may prove to be a considerable group of commercial photographs for which speedy distribution is of value.

Propagation of Electric Waves Over the Earth

By H. W. NICHOLS and J. C. SCHELLENG

SYNOPSIS. The comparatively poor transmission of radio waves of two or three hundred meters indicates some sort of selective effect in the atmosphere. Such an effect is found to result from the existence of free electrons in the atmosphere when the magnetic field of the earth is taken into account. In the earth's magnetic field, which is about one-half gauss, this selective effect will occur at a wave length of approximately 200 meters. Ionized hydrogen molecules or atoms result in resonant effects at frequencies of a few hundred cycles, this being outside of the radio range. The paper, however, takes into account the effects of ionized molecules as well as electrons.

The result of this combination is that the electric vector of a wave traveling parallel to the magnetic field is rotated. Waves traveling perpendicular to the magnetic field undergo double refraction. Critical effects are observed in rotation, bending of the wave and absorption at the resonant frequency. The paper develops the mathematical theory of these phenomena and gives formulas for the various effects to be expected.

THE problem of the propagation over the earth of electromagnetic waves such as are used in radio communication has attracted the attention of a number of investigators who have attacked the problem along somewhat different lines, with the purpose of offering an explanation of how electromagnetic waves can affect instruments at a great distance from the source in spite of the curvature of the earth. No attempt will be made here to describe adequately the various theories, but we remark that the theories of diffraction around a conducting sphere in otherwise empty space did not give satisfactory results and led to the necessity for the invention of a hypothetical conducting layer (Heaviside layer) whose aid is invoked to confine the wave between two concentric spherical shells. In many cases this Heaviside layer was considered to have the properties of a good conductor and it was supposed that a beam of short waves, for example, might be more or less regularly reflected back to the earth. The high conductivity of this layer was supposed to be due to the ionizing action of the sun or of particles invading the earth's atmosphere from outside and producing in the rarefied upper atmosphere a high degree of ionization. The differences in transmission during day and night and the variations which occur at sunrise and sunset were supposed to be due to the different ionizing effects of the sun's rays appropriate to the different times of day. The explanation of the phenomenon of "fading" or comparatively rapid fluctuations in the intensity of received signals could then be built up on the assumption of irregularities in the Heaviside layer producing either interference between waves arriving by different paths or reflection to different points on the earth's surface. The principal difficulty in this explanation is the necessity for rather high conductivity to account for the propagation of waves to great distances without large absorption.

In 1912 there appeared an article by Eccles [1] in which the bending of waves around the surface of the earth was explained on the basis of ions in the upper atmosphere which became more numerous as the vertical height increased and thereby decreased the effective dielectric constant which is a measure of the velocity of propagation of the wave. In this case the velocities at higher levels will be slightly greater than the velocities at lower levels, which will result in a bending downward of the wave normal and a consequent curvature of the wave path to conform to the curvature of the earth. In order to produce this effect without absorption the ions must be relatively free. If they suffer many collisions during the period of a wave, energy will be absorbed from the wave and pass into the thermal agitation of the molecules. Thus absorption of the wave can be computed provided the nature of the mechanism is understood thoroughly.

Sommerfeld and others have worked out the effect of the imperfect conductivity of the ground upon the wave front and such computations lead to a prediction that the electric vector in the wave near the ground will be tilted forward and thus have a horizontal component. This effect of imperfect conductivity is usually given as the cause of the large electromotive force which is induced in the so-called "wave antenna." This effect, however, apparently does not lead to an explanation of the bending of waves around the earth.

There has recently appeared an article by Larmor [2] in which the idea of a density gradient of ions or electrons is developed further to explain the bending of waves around the earth without a large absorption. This paper, as well as that of Eccles, leads to the conclusion that long radio waves will be bent around the earth, and that the effect increases as the square of the wave length, becoming vanishingly small for very short waves.

The large amount of data now available from both qualitative and quantitative observations of radio transmission shows that the phenomena may be more complicated than would be indicated by these theories. It is found that very long waves possess a considerable degree of stability and freedom from fading and that as the wave length decreases the attenuation and the magnitude of fluctuations increases until for a wave length of the order of two or three hundred

[1] Proc. Roy. Soc., June, 1912.

[2] *Phil. Mag.*, Dec., 1924.

meters there is great irregularity in transmission so that reliable communication over land for distances as short as 100 miles is not always possible even with large amounts of power. With decreasing wave length we find also variations in apparent direction of the wave. On the other hand, as the wave length is decreased still further we find, sometimes, rather surprising increases in range and stability. The nature of the fading changes, becoming more rapid, and the absorption in many cases seems to decrease. This peculiarity of wave transmission must be explained in a satisfactory theory. In addition to the apparent selective effect just mentioned, some observations indicate that there are often differences between east and west and north and south transmission at all wave lengths.

The various irregularities in radio transmission, and particularly the apparently erratic and anomalous behavior of electromagnetic waves occurring in the neighborhood of a few hundred meters wave length seem to indicate that as the wave length is decreased from a value of several kilometers to a value of a few meters some kind of selective effect occurs which changes the trend of the physical phenomena. These considerations have suggested to us the possibility of finding some selective mechanism in the earth's surface or in the atmosphere which becomes operative in the neighborhood of 200 meters. A rather superficial examination of the possibility that such a selective mechanism may be found in a possible distribution of charged particles in the atmosphere has resulted in the conclusion that a selective effect of the required kind cannot be produced by such a physical mechanism. There is, however, in the earth's atmosphere—in addition to distributions of ions—a magnetic field due to the earth, which in the presence of ions will have a disturbing effect upon an electromagnetic wave. As is well known, a free ion moving in a magnetic field has exerted upon it, due to the magnetic field, a force at right angles to its velocity and to the magnetic field. If the ion has impressed upon it a simple periodic electric force, it will execute a free oscillation together with a forced oscillation whose projection on a plane is an ellipse which is traversed in one period of the applied force. The component velocities are linear functions of the components of the electric field and at a certain frequency, depending only upon the magnetic field and the ratio $\frac{e}{m}$ of the ion, become very large unless limited by dissipation. This critical frequency is equal to $\frac{He}{2\pi mc}$ if H is measured in electromagnetic units and e in electrostatic units. It is the same as the frequency of free oscillation. For an electron in the earth's magnetic field (assumed to have a value of 1/2 gauss) this resonant frequency is 1.4×10^6 cycles, corresponding to a wave length of 214 meters.[3] We thus have an indication that some at least of the phenomena of transmission at the lower wave lengths may be explained by taking into account the action of the earth's magnetic field upon electrons present in the earth's atmosphere and acted upon by the electric field of the wave. This frequency occurs at approximately the position in the spectrum at which the peculiar effects already mentioned occur. The next resonant frequency which would be encountered would be due to the hydrogen ion which has a ratio, $\frac{e}{m}$, equal to $\frac{1}{1800}$ that of the electron. The resonant frequency of this ion is only 800 cycles and certainly can have no sharply selective effect in the propagation of electromagnetic waves over the earth. We have, therefore, worked out the consequences of the assumption that we have in the upper atmosphere two controlling factors influencing the propagation of electromagnetic waves in the radio range, namely, free electrons and ions together with the earth's magnetic field. The electrons will be dominant in their effects in the neighborhood of the resonant frequency and perhaps above, while the heavy ions will affect the wave at all frequencies and, if much more numerous, may be controlling at frequencies other than the critical one. In working out this theory it is assumed that there are present in the earth's atmosphere free electrons and ions. At high altitudes these are capable, on the average, of vibrating under the influence of the electromagnetic field through several complete oscillations before encountering other ions or neutral atoms. At low altitudes this assumption will not hold, the collisions being so numerous that the importance of the resistance term in the equations of motion becomes much greater. In either case the ions have no restoring forces of dielectric type. The motion of the electron or ion constitutes a convection current which reacts upon the electromagnetic wave and changes the velocity

[3] This frequency does not depend upon the direction of the field, and is practically constant over the earth's surface.

On March 7, after this paper had been written, the February 15 issue of the Proceedings of the Physical Society of London arrived in New York. In this journal there was a discussion on ionization in the atmosphere in which Prof. E. V. Appleton suggested, in an appendix, that the earth's magnetic field acting upon electrons would change the velocity of a wave and produce rotation. A calculation of the critical frequency was given in which, however, only the horizontal component of the earth's field was used, resulting in an incorrect value for the critical frequency, namely less than half the actual value. If the complete equations are written down it is evident at once that the total field is involved in the critical frequency, no matter what may be the direction of propagation.

of propagation of the wave. This is, in fact, the basis for the explanation of the optical properties of transparent and absorbing media and also of media which show magnetic or other rotatory powers. Due to collisions and recombinations, energy will pass continuously from the electromagnetic field and increase the energy of agitation of neutral molecules. Since this process is irreversible it accounts for absorption of energy from the wave.

Assume an electron or ion of charge e and mass m moving with velocity $\mathbf{v}$ and acted upon by an electric field $\boldsymbol{E}$ and the earth's magnetic field $\boldsymbol{H}$. The equation of motion of the free ion will be

$$\frac{m}{e}\dot{\mathbf{v}}=\boldsymbol{E}+\frac{1}{c}\mathbf{v}\times\boldsymbol{H}$$

or

$$a\dot{\mathbf{v}}=\boldsymbol{E}+\mathbf{v}\times\boldsymbol{h} \tag{1}$$

in which $\boldsymbol{h}$ is written for $\frac{\boldsymbol{H}}{c}$ and a for m/e. (When we come to consider absorption it will be necessary to generalize a into $a\left(1-i\,\frac{r}{mn}\right)$ to include a resisting force, rv, proportional to the velocity.)

The total current is given by

$$4\pi\boldsymbol{I}=\dot{\boldsymbol{E}}+\sum 4\pi Ne\,\mathbf{v}. \tag{2}$$

In these equations and the following we are using Gaussian units and the summation refers to different kinds of ions.

In order to avoid a complicated mathematical treatment, which, however, is not difficult to carry through if necessary, it will be assumed that the magnetic field $\boldsymbol{H}$ is along the axis of z. When more general results are required, they will be stated. All time variables are assumed periodic with a frequency $\frac{n}{2\pi}$, so that $\frac{\partial}{\partial t}=in$.

Solving equation (1) for the components of $\mathbf{v}$ we find, for each type of ion:

$$v_1=\frac{ina\,X+h\,Y}{h^2-a^2n^2},$$

$$v_2=\frac{-hX+ina\,Y}{h^2-a^2n^2},$$

$$v_3=\frac{Z}{ina},$$

from which it appears that a resonance frequency occurs for

$$n = \frac{h}{a} = n_o.$$

Since e/m for the electron is $-1.77c \times 10^7$, the earth's magnetic field of about 1/2 gauss will produce a resonance frequency at 1.4×10^6 corresponding to a wave length of 214 meters, while all heavier ions have resonance frequencies far outside the spectral region to be considered.

The assumption that the components of the ionic motion are simple harmonic, in spite of the fact that the motion of the ion is rather complicated, is justified as follows. From (1) we find that the velocity of an ion (r), say $\mathbf{v}_r$ is made up of the complementary solution, $\mathbf{v}_r'$ and the particular solution $\mathbf{v}_r'' = f(\boldsymbol{E})$. The latter depends upon the impressed force $\boldsymbol{E}$, while the former has constants of integration determined by the position and motion of the ion at the last collision. The complete current is thus

$$\boldsymbol{I} = \frac{1}{4\pi}\dot{\boldsymbol{E}} + \sum e\mathbf{v}_r' + Nef(\boldsymbol{E}).$$

The second term, however, averages out over a large number of ions since the initial conditions are random;[4] hence, as far as the effect upon wave propagation is concerned, we may treat all quantities as periodic.

Following the usual procedure for the investigation of the propagation of waves in media of this kind, we shall rewrite equation (2) in terms of the components of the electric field, thus for each type of ion:

$$4\pi I_1 = \left(1 + \frac{\sigma N}{n_o^2 - n^2}\right)\dot{X} - i\,\frac{\sigma N \frac{n_o}{n}}{n_o^2 - n^2}\dot{Y} = \epsilon_1\dot{X} - i\alpha\dot{Y},$$

$$4\pi I_2 = i\,\frac{\sigma N \frac{n_o}{n}}{n_o^2 - n^2}\dot{X} + \left(1 + \frac{\sigma N}{n_o^2 - n^2}\right)\dot{Y} = i\alpha\dot{X} + \epsilon_1\dot{Y}, \qquad (3)$$

$$4\pi I_3 = \left(1 - \frac{\sigma N}{n^2}\right)\dot{Z} \qquad = \epsilon_2\dot{Z},$$

in which $\frac{4\pi e}{a} = \sigma$, or 3.2×10^9 for an electron and $3.2 \cdot 10^9 \frac{m}{M}$ for an ion of mass M. In order to avoid complicated formulas, the summations which must be carried in equations (3) to take account

[4] It is here assumed that the mean time between collisions is large compared to $\frac{1}{n}$.

of the effect of ions of different kinds have been omitted, but it is to be understood that the dielectric constants ϵ, α, etc., are built up from the contributions of all types of ions. Thus for an ion of mass M we must put $\sigma \frac{m}{M}$ for σ, $n_0 \frac{m}{M}$ for n_0, in equations (3).

The effective dielectric constant, instead of being unity, has thus the structure:

$$(\epsilon) = \begin{pmatrix} \epsilon_1 & -i\alpha & 0 \\ i\alpha & \epsilon_1 & 0 \\ 0 & 0 & \epsilon_2 \end{pmatrix}$$

and we may write equation (2) as

$$4\pi \boldsymbol{I} = (\epsilon)\dot{\boldsymbol{E}}$$

which has the significance of the scalar equations (3). Thus $\boldsymbol{I}$ is a linear vector function of $\boldsymbol{E}$ and the operator (ϵ) is skew symmetric, indicating a rotatory effect about the axis of z.

(The general case in which h has the three components $(h_1\ h_2\ h_3)$ results in a dielectric constant having the structure

$$(\epsilon) = \begin{pmatrix} \epsilon_1 & -\beta_3 - i\alpha_3 & -\beta_2 + i\alpha_2 \\ -\beta_3 + i\alpha_3 & \epsilon_2 & -\beta_1 - i\alpha_1 \\ -\beta_2 - i\alpha_2 & -\beta_1 + i\alpha_1 & \epsilon_3 \end{pmatrix}$$

of which the above is a special case. With this value of (ϵ) the equation (4) below contains the *general* solution of our problem.)

Let $\boldsymbol{H}_1$ be the magnetic force associated with $\boldsymbol{E}$ in the wave so that

$$c \text{ curl } \boldsymbol{H}_1 = (\epsilon)\ \dot{\boldsymbol{E}}$$

$$c \text{ curl } \boldsymbol{E} = -\dot{\boldsymbol{H}}_1.$$

Eliminating $\boldsymbol{H}_1$ from these equations we get

$$-\nabla^2 \boldsymbol{E} + \nabla \text{ div } \boldsymbol{E} = \frac{n^2}{c^2}(\epsilon)\boldsymbol{E} \tag{4}$$

or in scalar form

$$-\nabla^2 X + \frac{\partial}{\partial x} \text{ div } \boldsymbol{E} = \frac{n^2}{c^2}(\epsilon_1 X - i\alpha Y),$$

$$-\nabla^2 Y + \frac{\partial}{\partial y} \operatorname{div} E = \frac{n^2}{c^2}(i\alpha X + \epsilon_1 Y), \tag{5}$$

$$-\nabla^2 Z + \frac{\partial}{\partial z} \operatorname{div} E = \frac{n^2}{c^2}(\epsilon_2 Z).$$

These equations for the propagation of light in magnetically active substances have been given by Voigt, Lorentz, Drude and others and form the basis of the explanation of optical phenomena in such substances. As applied to optics, they are worked out, for example, in Drude's "Optics" (English translation), page 433. As applied to this problem, they assume either that the motion of the ions is unimpeded or that the resistance to the motion may be expressed as a constant times the velocity, which, as explained later, may be done in this case. We shall work out some comparatively simple cases and point out the conclusions to be drawn from them.

Consider first a plane polarized ray having its electric vector parallel to the magnetic field and moving in the xy plane; for example parallel to x. In this case the electric vector is a function of x and t only of the form

$$Z = Z_o \epsilon^{in\left(t - \frac{\mu x}{c}\right)}$$

in which $\frac{c}{\mu}$ is the velocity of the wave. Substituting in the general equations (5) we find that

$$\mu^2 = 1 - \Sigma \frac{\sigma_i N_i}{n^2}. \tag{6}$$

The velocity of propagation is thus a function of the frequency and of the density N. This particular case corresponds to that treated by Eccles and Larmor in the papers cited. It will be noted that the velocity is greater for long waves than for short waves and that if N is a function of distance from the surface of the earth, the velocity will vary in a vertical direction, causing a curvature of the rays as worked out by the authors mentioned. In this particular case, however, which corresponds completely in practice to conditions obtaining over only a limited area of the earth's surface, the greatest effect is produced on the longer waves. Since electromagnetic waves are in general radiated from vertical antennas so that the electric vector is vertical, this case would correspond to the condition of transmitting across the north or south magnetic poles of the earth.

The second case to be considered is that of propagation along the direction of the magnetic field. In this case X and Y are functions of z and t and the appropriate solutions of the fundamental equations (5) are

$$X' = A \cos n\left(t - \frac{\mu_1 z}{c}\right),$$

$$Y' = -A \sin n\left(t - \frac{\mu_1 z}{c}\right), \quad \mu_1^2 = \epsilon_1 + \alpha,$$

$$X'' = A \cos n\left(t - \frac{\mu_2 z}{c}\right),$$

$$Y'' = A \sin n\left(t - \frac{\mu_2 z}{c}\right), \quad \mu_2^2 = \epsilon_1 - \alpha.$$

which represent two oppositely circularly polarized components traveling with the different velocities $\frac{c}{\mu_1}$ and $\frac{c}{\mu_2}$. The plane of polarization is rotated through an angle of 2π in a distance given by

$$\frac{z_0}{\lambda} = \frac{\epsilon_1}{\alpha}.$$

The third case to be considered is that of propagation at right angles to the magnetic field, say in the direction of x. For this case equations (5) become:

$$X = \frac{i\alpha}{\epsilon_1} Y$$

$$-\frac{c^2}{n^2}\nabla^2 Y = \left(\epsilon_1 - \frac{\alpha^2}{\epsilon_1}\right) Y$$

$$-\frac{c^2}{n^2}\nabla^2 Z = \epsilon_2 Z,$$

of which the solutions are

$$X = \frac{i\alpha}{\epsilon_1} Y_0 \epsilon^{in\left(t - \frac{\mu_1 x}{c}\right)}$$

$$Y = Y_0 \epsilon^{in\left(t - \frac{\mu_1 x}{c}\right)} \quad \mu_1^2 = \epsilon_1 - \frac{\alpha^2}{\epsilon_1}.$$

$$Z = Z_0 \epsilon^{in\left(t - \frac{\mu_2 x}{c}\right)}.$$

The first of these is merely the (usually small) component of field required to make the total current solenoidal, that is, to balance the convection of electrons. The last two show that the plane polarized ray whose electric vector is parallel to H will travel with the velocity $\frac{c}{\mu_2}$ while the one whose electric vector is at right angles to this direction and to the direction of propagation will travel at a different speed, $\frac{c}{\mu_1}$. There is thus double refraction.

Bending of the rays. If μ is the index of refraction, which is a function of the space variables, the curvature of the ray having this index is $\frac{1}{\mu}\frac{d\mu}{ds}$ where s is taken perpendicular to the direction of the ray. Since μ is practically unity except at the critical frequency, this curvature is $1/2\, d\mu^2/ds$. In order that the ray should follow the curvature of the earth it is clear that μ must decrease at higher altitudes; that is, $\frac{d\mu^2}{ds}$ must be negative.

We shall work out the curvatures for the special cases considered. (The first case has been given above and was worked out in the papers cited). For the case of propagation along H, the two circularly polarized beams have indices given by

$$\mu_1^2 = \epsilon_1 + \alpha = 1 + \frac{\sigma N}{n^2}\,\frac{1}{\omega - 1}, \tag{7}$$

$$\mu_2^2 = \epsilon_1 - \alpha = 1 - \frac{\sigma N}{n^2}\,\frac{1}{\omega + 1}, \tag{8}$$

$$\left(\omega = \frac{n_0}{n}\right).$$

We are interested in the values of $1/2\,\frac{d\mu^2}{ds}$ in which N and h are functions of distance s and also of the time. These come out to be

$$C_1 = \frac{\sigma}{2n_0^2}\left[\frac{\omega^2}{\omega - 1}\,\frac{dN}{ds} - \frac{\omega^3}{(\omega - 1)^2}\,\frac{N}{h}\,\frac{dh}{ds}\right], \tag{9}$$

$$C_2 = \frac{\sigma}{2n_0^2}\left[\frac{-\omega^2}{\omega + 1}\,\frac{dN}{ds} + \frac{\omega^3}{(\omega + 1)^2}\,\frac{N}{h}\,\frac{dh}{ds}\right]. \tag{10}$$

A striking fact shown by these formulae is that the curvatures of the two rays are in general different. A limited beam entering an ionized medium along a magnetic meridian will be split into two which will traverse different paths. Thus we should expect to find, occasionally, a circularly polarized beam at the receiver due to the fact that the receiving instrument is located at a point toward which one of the beams is diverted after having passed through an upper ionized layer. This is now being investigated experimentally. It is clear that, although the two components do not in general travel over the same path, both may eventually arrive at the same receiver. The first ray, however, may have penetrated much higher in the atmosphere than the other, that is, to a level at which $\frac{dN}{ds}$ has the proper negative value to cause it to return to earth.

For long waves, these curvatures become:

$$C_1 = \frac{\sigma\omega}{2n_o^2}\left[+\frac{dN}{ds} - \frac{N}{h}\frac{dh}{ds}\right], \tag{11}$$

$$C_2 = \frac{\sigma\omega}{2n_o^2}\left[-\frac{dN}{ds} + \frac{N}{h}\frac{dh}{ds}\right]. \tag{12}$$

Hence a limited beam of long waves entering this medium would tend to split into two of opposite polarization and traverse different paths. In the special case for which $\frac{1}{N}\frac{dN}{ds} = \frac{1}{h}\frac{dh}{ds}$ throughout the medium, there will be no such separation of the beam.

For very short waves

$$C_1 = \frac{\sigma}{2n_o^2}\left[-\omega^2\frac{dN}{ds} - \omega^3\frac{N}{h}\frac{dh}{ds}\right], \tag{13}$$

$$C_2 = \frac{\sigma}{2n_o^2}\left[-\omega^2\frac{dN}{ds} + \omega^3\frac{N}{h}\frac{dh}{ds}\right]. \tag{14}$$

Hence if the most effective cause of refraction is the variation in the ionic density both components tend to remain together and to travel with a rotation of the plane of polarization. If variation in the magnetic field is appreciable the two components tend to diverge as in the case of long waves.

For propagation at right angles to H, say along x, we have

$$\mu_1^2 = \epsilon_2 = 1 - \frac{\sigma N}{n^2}, \tag{15}$$

$$\mu_2^2 = \epsilon_1 - \frac{\alpha^2}{\epsilon_1}. \tag{16}$$

The bending of the plane polarized component having the index μ_1 shows no selective effects, being simply

$$C_1 = -\frac{\sigma}{2n^2}\frac{dN}{ds} \tag{17}$$

and is appreciable only for long waves unless N is very large. For the other component we find:

$$C_2 = \frac{\sigma}{2n_o^2}\cdot\frac{\omega^2}{\omega^2-1}\cdot\frac{1-\dfrac{2\sigma N}{n_o^2}\omega^2-\dfrac{\sigma^2 N^2}{n_o^4}\dfrac{\omega^2}{\omega^2-1}}{\left(1+\dfrac{\sigma N}{n_o^2}\dfrac{\omega^2}{\omega^2-1}\right)^2}\frac{dN}{ds} \tag{18}$$

where, in order to simplify the formula, only the term containing $\frac{dN}{ds}$ has been included. This applies to ions of one kind.

For long waves these two curvatures become

$$C_1 = -\frac{\sigma}{2n_o^2}\omega^2\frac{dN}{ds}, \tag{19}$$

$$C_2 = \frac{\sigma}{2n_o^2}\left(1-\frac{2\sigma N}{n_o^2}\omega^2\right)\frac{dN}{ds}. \tag{20}$$

These formulas show that the first curvature is always in the same direction for a given value of $\frac{dN}{ds}$, while the second curvature, which is that of the electric vector perpendicular to the magnetic field, is, for very long waves, in the same direction as C_1 but, as the wave length is decreased or N increased, reverses in sign and becomes opposite to C_1. As an example, if $N=10$, for 6 kilometer waves the curvatures are opposite, so that if the first component tends to bend downward the second will tend to bend upward; while if $N=100$, for the same wave length both curvatures have the same sign and the second is five times as large as the first.

For extremely short waves the two curvatures are equal as they obviously should be, since the magnetic field can then have no effect.

In transmitting from New York to London, for example, waves travel approximately at right angles to the magnetic field, which in this latitude has a dip of about 70°. If we assume a plane polarized ray starting out with its electric vector vertical, the component parallel to the magnetic field will be the larger and will be subject to the curvature C_1 above, while the smaller component will be affected by the magnetic field and will have the curvature C_2. The two components into which the original wave is resolved will travel with different velocities. It is clear that when the distribution of ions in the upper atmosphere is changed by varying sunlight conditions, the resulting effect at a receiver is likely to vary considerably. Some of the possibilities will be discussed later.

Rotation of the plane of polarization. It has been shown that in the second case, namely transmission along the magnetic field, there will be a rotation of the plane of polarization of the wave. This rotation is such that the wave is rotated through a complete turn in a distance given by

$$z = \frac{2\pi c}{n_o} \frac{1 + \frac{\sigma N}{n_o^2} \frac{\omega^2}{\omega^2 - 1}}{\frac{\sigma N}{n_o^2} \frac{\omega^2}{\omega^2 - 1}}. \tag{21}$$

It is interesting to note that the distance in which a long wave rotates through 2π approaches the constant value $\frac{2\pi c n_o}{\sigma N}$ as the wave length increases and that for very short waves the rotation of the plane of polarization tends to vanish with the wave length.

Absorption. When an electron strikes a massive neutral atom the average persistence of velocities is negligible and in the steady state of motion of electrons and neutral molecules the element of convection current represented by an impinging electron will be neutralized, so far as the wave is concerned, at every collision. Of the energy which has been put into this element of convection current since the last collision, a part will be spent in accelerating neutral molecules, part will go to increase the average random velocity of the electron and a part will appear as *disordered* electromagnetic radiation. Thus, as far as the wave is concerned, the process of collision with massive neutral molecules is irreversible even if the molecules are elastic, and all the energy picked up by the electron from the wave between collisions is taken from the wave at the next collision. Exactly the same state of affairs would exist if at each collision the electron recombined with a molecule and a new electron were created with zero or random velocity. Thus for massive molecules for which we can neglect the persistence of electron velocities the effect upon the wave is exactly the same whether the collision is elastic or inelastic.

These conclusions are verified by the results of two different computations which we have made of the resistance term, rv, in equation of motion of the electron. Consider in the first place a mixture of electrons and massive neutral molecules, assumed perfectly elastic, in which the persistence of velocities of the electrons after collision is negligible. If an electric field $X\epsilon^{int}$ operates in the x direction and if the state of motion is a steady one, we can compute the energy w taken from the wave by a single electron at any time after a collision at the time t_1 and before the next collision. Let this time after t_1 be τ. If the mean frequency of collisions is f, the time τ between collisions will be distributed according to the law

$$f\epsilon^{-f\tau}$$

and we shall obtain the mean energy taken from the wave per collision by multiplying w by the above expression, integrating from zero to infinity with respect to τ and then performing an average over all the times t_1. The result of this is that the mean energy loss per collision is simply

$$w=\frac{e^2X^2}{2mn^2}\,\frac{n^2}{f^2+n^2}$$

and consequently the loss per second is f times this. If we equate this to rv^2, which is also the rate at which energy is being dissipated, we find that $r=mf$, which is therefore the resistance term to be inserted in the equation of motion of the electron.

If the convection current is carried partly by heavier ions, it will not be annulled at each collision and all the energy derived from the field will not be lost on impact.

The foregoing computation assumes as obvious that energy is lost from the wave at a rate equal to the number of collisions times the average energy which the electron takes from the wave between collisions. The second method is somewhat more general. The mean velocity at a time t is found for electrons which collided last in an interval at t_1. This is evidently a function of the velocity persisting through the last collision and hence of the average velocity before the impact; so that if the average velocity before collision was v, that after impact would be δv, in which δ is a number less than unity, depending on the relative masses and the nature of the collision. Averaging for all values of t_1 before t and using the same law of distribution assumed above, the mean velocity of the ions since the last collision is obtained. By comparison with the solution obtained for the velocity of forced oscillation in which the resistive force is rv, we find that $r=mf(1-\delta)$. For the special case of electrons, δ may be taken equal to zero, hence $r=mf$. For the case of very heavy ions colliding with light neutral molecules, $r = 0$, since $\delta = 1$. For equal masses δ would be about one half, hence $r = \frac{1}{2}\, mf$.

Since the resistance factor r is equal to mf, in order to include the effect of attenuation of the wave, we must replace a by

$$a\left(1 - i\frac{f}{n}\right).$$

If, as usual, we assume a wave proportional to

$$\epsilon^{-\frac{nk\mu x}{c}} \quad \epsilon^{in\left(t - \frac{\mu x}{c}\right)}$$

the equations (5) show that, in order to calculate the value of the absorption constant k, we must put

$$\mu^2(1 - ik)^2 = \epsilon$$

in which ϵ is the generalized dielectric constant appropriate to the case considered. We have worked out in this way the absorption for the various cases treated above with the following results.

In the case in which there is either no magnetic field or the magnetic field is parallel to the direction of the electric vector, we find

$$k = \frac{\sigma N}{2 n_0^2}\, \omega^2 \quad \frac{f/n}{1 + f^2/n^2}.$$

This formula for absorption applies (for electrons) for any value of f or n. Thus near the surface of the earth where the collision frequency f is of the order of 10^9, the fraction $\frac{f}{n}$ will be large even for rather short waves. As we go higher in the atmosphere this ratio decreases for a given wave frequency until at a height for which $\frac{f}{n} = 1$ we encounter the maximum absorption *per electron*. Above this level $\frac{f}{n}$ and consequently the absorption per electron decreases. For ions other than electrons the resistance will be somewhat different from mf, depending upon the ratio of the masses, and a corresponding change must be made in the above statement.

In this paper we are considering only the effects which take place at heights above that for maximum absorption so that, generally speaking, $\frac{f}{n}$ will be small or at least less than unity. This approximation will be used in computing the absorption constants which follow.

As an example of the nature of this approximation, at a height of about 100 kilometers, we may expect an atmospheric pressure of 10^{-5} standard and a corresponding collision frequency of the order of 10^5. Thus for very long waves of frequency 40,000 cycles per second we still have $\frac{f}{n} = .4$, while at the critical frequency $\frac{f}{n}$ is only 1/100.

The computation of the collision frequency for electrons is rather involved because of the peculiar nature which such a collision may have and because it probably is not permissible to assume thermal equilibrium with the molecules of the gas. The processes of ionization and recombination will also lead to complications. Probably the most significant information would be the number of electron free paths per second for unit volume.

The question of the behavior of waves in or below the layer of maximum absorption per ion is a somewhat different one and belongs properly in another paper.

For the case of transmission along a magnetic meridian the oppositely circularly polarized rays have the absorption constants:

$$k_1 = \frac{\sigma N}{2n_o^2} \frac{\omega^2 f/n}{(\omega-1)^2+(f/n)^2}, \qquad k_2 = \frac{\sigma N}{2n_o^2} \frac{\omega^2}{(\omega+1)^2} \frac{f}{n}.$$

It will be noted that, at the critical frequency, the first of these waves has the high absorption $\frac{\sigma N}{2n_o^2} \cdot \frac{n}{f}$ and is therefore extinguished in a short distance, while the other wave has a normal absorption constant $\frac{\sigma N}{8n_o^2} \cdot \frac{f}{n}$. Thus for the case of transmission along a meridian at the critical frequency we might expect a receiving station, sufficiently far above the ground, to receive a circularly polarized beam. This would mean that if a loop were used for reception, the intensity of the received signal would be independent of the angle of setting of the loop, provided one diameter of the loop was set parallel to the direction of propagation of the wave. In general, of course, this ideal condition could not be realized due to the disturbing action of the ground and of other conducting or refracting bodies and the most we should expect to receive in practice would be an elliptically polarized beam.

In the third case, namely, that of propagation perpendicular to the direction of the magnetic field, we find that the wave polarized with its electric vector parallel to the magnetic field has the same absorption as before, namely $\frac{2n_o^2}{\sigma N}\cdot\omega^2\frac{f}{n}$ and the other ray whose complex index of refraction is $\epsilon_1 - \frac{\alpha^2}{\epsilon_1}$ has the absorption constant $\frac{1}{2}(k_1+k_2)$ in which k_1 and k_2 are the absorption constants given above for propagation along a magnetic meridian.

At the critical frequency we find, therefore, that the absorption constant is abnormally high and equal to $\frac{\sigma N}{4n_o^2}\cdot\frac{n}{f}$ which is one-half that obtained for the first ray of case 2.

One very striking fact is brought to light by these equations. Thus, referring to the two values of absorption constants for transmission along the magnetic field, we find that for very long waves (for which ω is large) the ionic absorption is very much less with a magnetic field present than without it. This means that in this case and in the next the presence of a magnetic field *assists* in the propagation of an electromagnetic wave by decreasing the absorption. This reduction in absorption may amount to a rather large amount, as may be seen from an inspection of the formula for k_1. For example, if in this case ω is 20, corresponding to 4,000 meter waves, we find that under corresponding conditions the absorption *due to electrons only* is reduced by the magnetic field to 1/400th the value it would have for no magnetic field. Of course, these cases are not directly comparable because the path chosen by the wave would be different in the two cases. It is plausible, however, that the propagation of long waves along the magnetic field may go on with much less attenuation than propagation from East to West over a region in which the magnetic field is nearly vertical, in which case the effect of the magnetic field is largely absent. This conclusion, however, cannot be made in general since a number of other causes are influential in determining the propagation, for example, the bending of the rays, so that it is not certain that transmission over a region in which the magnetic field is vertical is always more difficult than in the other cases.

The reason for the decreased absorption of long waves when the magnetic field can operate (that is, in all cases in which the electric vector is not parallel to the field) is that the velocities acquired by the free electrons are much less for small values of n when the magnetic field is present.

Fading. By this is meant a variation with time of the strength of a received signal at a given point. It is clear that a wave starting originally with constant amplitude and frequency can be received as one of variable amplitude only if certain characteristics of the medium are variable with the time. So far as the atmosphere is concerned, these characteristics may be the distribution of electrons and heavier ions and the intensity and direction of the earth's magnetic field. If these are functions of the time, the velocities, bending, absorption and rotation of the plane of polarization will all be variable, the amplitude of variation depending upon the variations of N, $\frac{dN}{ds}$, H, $\frac{dH}{ds}$, as well as the frequency of the wave, the effects being in many cases magnified greatly in the neighborhood of the critical frequency. These effects are obviously sufficiently numerous to account for fading of almost any character and suggest a number of experiments to determine the most effective causes. The question of rotation of the plane of polarization, fading and distortion is now being examined experimentally.

From the formulas it is clear that the velocity, curvature and absorption of an electromagnetic wave as well as the rotation of its plane of polarization can all be affected by a time variation in the intensity and direction of the earth's field. An examination of the probable time and space variations of each, however, lead us to the conclusion that these are not of primary importance in determining large amplitude fading except, perhaps, during magnetic storms. One result of the last two years of consistent testing between New York and London at about 60,000 cycles has shown that severe magnetic storms are always accompanied by corresponding variations in the strength of received signals. Thus, although the earth's magnetic field can well exercise a large influence upon the course and attenuation of radio waves, it does not seem likely that its time variation is ordinarily a large contributing cause to fading.

This leaves as the probable principal cause of time variations the number and distribution of ions in the earth's atmosphere. It is impossible in this paper, which is devoted primarily to a development of a theory of transmission involving the earth's magnetic field, to consider adequately all the possibilities resulting from changes in ionic distributions, but some general remarks may be made. Imagine a wave traveling from the source to the receiver. At a short distance from the source the wave front will be more or less regular but as it progresses, due to the irregularities in ionic distribution, the wave front will develop crinkles which become exaggerated as the wave goes on. These crinkles in the wave front will be due to irregularities in the medium and can be obtained by a Huyghen's construction at any point. If we consider the wave a short distance before it reaches the receiver, we will find regions in which the wave front is concave to the receiver and regions of opposite curvature. Thus at certain portions of the wave front energy will be concentrated toward a point farther on and at other parts will be scattered. The location of these convex or concave portions of the wave in the neighborhood of a given receiving point will be very sensitive to changes in ionic distribution along all the paths of the elementary rays contributing to the effect at the receiver. Hence, if we knew the location and movement of all the ions between the transmitter and the receiver, it would be possible, theoretically, to predict the resultant effect at the latter point.

To explain fading it is essential that there be a time variation in this distribution. It is clear that effects of this kind should be more marked at short waves than at long waves since a region of the medium comparable in dimensions to a wave length must suffer some change in order to produce an effect upon the received signal. If, for example, there were space irregularities in the medium comparable to the wave length, a kind of diffraction effect would be produced at the receiver which would be very sensitive to slight changes in grating space.

A possible cause of irregularity may be found in the passage across the atmosphere of long waves of condensation and rarefaction, each of which results in a change in the density and gradient of the ions, even though the average density remains constant throughout a large volume. If, as seems plausible, the upper atmosphere is traversed by many such atmospheric waves of great wave length, the resulting effect at a given receiving point would be fluctuations in signal strength due to a more or less rapid change in the configuration of the wave front near the receiver.

For radio waves whose length is of the order of a few hundred meters, fading experimentally observed occurs at a rate of the order of one per minute (of course, it is not implied by this statement that there is any regular periodicity to the fading). The pressure wave referred to would travel in the upper atmosphere with a velocity of the order of 300 meters per second at lower levels or 1,000 meters in the hydrogen atmosphere, so that the wave length of these "sound" waves would be of the order of 50 of the radio wave lengths. The irregularities of the medium would thus be of sufficient dimensions with respect to the electromagnetic waves so that one of the characteristics referred to above might be developed. In this way we might explain variations in intensity of the wave at the receiver recurring at intervals of a minute or so.

These effects, of course, might be produced even without a magnetic field but the results of this paper indicate that conditions in the wave front will be complicated still further by a rotation of the electric vector and by the existence of bending and double refraction due to the magnetic field, these effects being exaggerated in the neighborhood of the critical frequency. Due to the magnetic field we have also the possibility of summation effects between components of the wave which were split off by the action of the field and consequently had traveled by different paths at different speeds. It is obviously impossible to make any general statement concerning the nature of the effects which will be produced by this complicated array of causes but future experimental work will, we hope, allow us to estimate the relative importance of the various elements.

Open Tank Creosoting Plants for Treating Chestnut Poles

By T. C. SMITH

Introduction

FOR a number of years chestnut timber, because of its many desirable characteristics, has served a broad field of usefulness in telephone line construction work, not only in its native territory, the eastern and southeastern part of the United States, but also in neighboring states. In fact, as an average, about 200,000 chestnut poles are set annually in the Bell System plant as replacements and in new lines.

In areas which are gradually being extended from the northern part of the chestnut growing territory into the southern sections, blight is rapidly making serious inroads into this class of pole timber. North of the Potomac River practically all chestnut territories have been visited by the blight and it has in a major sense crossed into areas south and southwest of this river, where it is developing from scattered spots. While many poles are yet secured in the blighted areas, they must be cut within a very few years after becoming affected, in order to save them from the decay which destroys blighted poles after they are killed.

A chestnut pole lasts satisfactorily above the ground line but decays at and within a few inches below the ground, thus weakening it at a critical location. In order to protect the poles from decay at this location, the open tank creosote treatment seems to be the most satisfactory, where the facilities for applying the treatment are available. In general this treatment consists of standing the poles in an open tank and treating them in a creosote bath which covers them from the butt ends to a point about one foot above what will be the ground line when the poles are set. The method of applying the treatment will be explained in more detail further along in the paper.

Due to the scattered locations of the chestnut timber and also to the fact that in many places this timber is rapidly being depleted by the blight, it has required considerable study to establish locations for open tank treating plants which would be convenient for applying the treatments and would also have a sufficient available pole supply to permit the operation of the plants long enough to warrant the necessary investment in them. However, suitable locations have been established and plants have been constructed which will, when operating to their planned capacities, treat about 139,000 chestnut poles per year, and these plants may easily be enlarged to treat additional quantities as the demand for treated poles develops.

These plants have been designed by our engineers and are being operated for applying preservative treatments to poles used by the Bell System.

Locating the Treating Plants

It might be interesting to bring out the governing considerations in locating the chestnut open tank treating plants, as compared with commercial plants for treating cedar poles, which are operating in the north central and northwest portions of the United States. Due to the geographical locations in which the cedar poles grow, in relation to the centers of distribution en route to the locations where they will be used, treating plants of large capacities can be supplied for many years with poles which pass them in the normal course of transporting the poles from the timber to their destinations. Commercial pole treating companies seem to have had no difficulty in establishing locations for handling 100,000 or more cedar poles per year through a single plant; whereas the scattered locations of the chestnut poles, as outlined above, make it more economical to build the chestnut treating plants in units varying between 10,000 and 36,000 poles per year capacity.

Several factors were considered in determining the proper locations for the seven Bell System treating plants which have been built. It was often possible to select a location which was admirably adapted to the purpose when considered from two or three viewpoints but which was found undesirable when considered from all of the necessary angles. The principal points considered were:

1. Quantity of poles of the desired sizes available locally which could be delivered to a proposed plant by wagons, motor vehicles, etc.

2. Quantity of poles which could be conveniently routed past the plant during the rail shipments from the timber to their destinations.

3. Quality of the available timber.

4. The length of time during which a plant of the desired size could be supplied with timber for treatment. This estimated figure would, of course, determine the length of life of the proposed plant.

5. Railroad facilities and freight distances from the proposed plant to points where the poles would be used.

6. Availability of labor for operating the plant.

7. Locating a suitable site for the plant.

Experience of the Western Electric Company's Purchasing Department and the local Associated Telephone Company representatives, together with information from Government reports, provided the

Fig. 1—Land upon which Sylva Plant was Built

answers to the first five items. Studies upon the ground were made to settle the remaining two items after a preliminary survey of the situation had indicated what locations seemed to warrant consideration.

The unevenness of the land as shown by Fig. 1, which is typical of the many available locations studied, made it difficult to secure a comparatively level tract of the proper area and dimensions adjoining a railroad siding or at a location where a siding could conveniently be established. In fact it soon became evident in making the preliminary studies, that it would be necessary to design the various treating plants to fit the best of the available tracts.

As a result of these studies, seven plants were established and placed in operation in five states as outlined below:

Location	Date when Plant Was Placed in Operation	Annual Pole Capacity Now	Total Annual Pole Capacity When Additions Now Planned Are Completed
Shipman, Va.	Oct. 1922	10,000	15,000
Danbury, Conn.	Dec. 1922	10,000	10,000
Natural Bridge, Va.	Apr. 1923	10,000	18,000
Willimantic, Conn.	Aug. 1923	10,000	10,000
Sylva, N. C.	May 1924	18,000	25,000
Nashville, Tenn.	July 1924	18,000	25,000
Ceredo, W. Va.	Sept. 1924	23,000	36,000
Totals		99,000	139,000

It will be noted from the above table that several of the plants are not yet working to their capacities as now planned. In designing the plants, the plans were made to provide for the total annual capacities shown above. However, when they were built the initial capacities were made somewhat lower as indicated by the table, by omitting in some cases tanks and in other cases pole handling equipment which could readily be added in conformity with the plans, later when the additional capacities would be required.

Yard Sizes

It might not seem necessary to occupy a very great area in the operation of a pole treating plant. However, experience with some of the earlier plants indicated that a reasonably large yard was very desirable because of the number of poles necessarily carried in piles on skids in the yard both in the untreated stock and in the treated stock. In so far as practicable the poles in the various treating plants are arranged in such a manner that each length and class is piled separately. This greatly facilitates handling the poles, but requires considerable space. Ordinarily about 80 pole piles are necessary in a yard.

From four to ten acres of land has been used for each of the various pole treating yards. Fig. 2, which includes about half of a comparatively small capacity yard, shows the necessity for plenty of room for the pole piles.

Yard Layouts

Since the pole treating yard layouts are necessarily built around the railroad sidings which handle the poles in and out of the yards and transfer them from one location to another inside the yards, it is desirable to build the yards long and narrow.

Fig. 2—Portion of Pole Yard at One of the Smaller Plants. Tool House and Creosote Storage Tank at Right

Of course, the sharper the railroad curves can be made in laying out a siding from the railroad into the pole treating yard, the easier it is to accommodate the siding to cramped yard conditions or to spread out the tracks over a short, wide yard. However, due to the use of heavy locomotives on the main lines and the desirability of having switch curves suitable for the locomotives ordinarily used, it has been necessary to use 12 degree railroad curves in planning most of the yard entrances, and in no case has a curve been used which is sharper than 18 degrees.

It will be noted from Fig. 3 that the pole treating apparatus is so located that the work of handling poles to and from the treating tanks will not interfere in any way with loading outgoing cars of treated poles from the skids. It will also be noted that the poles which are received from the river are treated during the natural course of their passage to the "treated" skids.

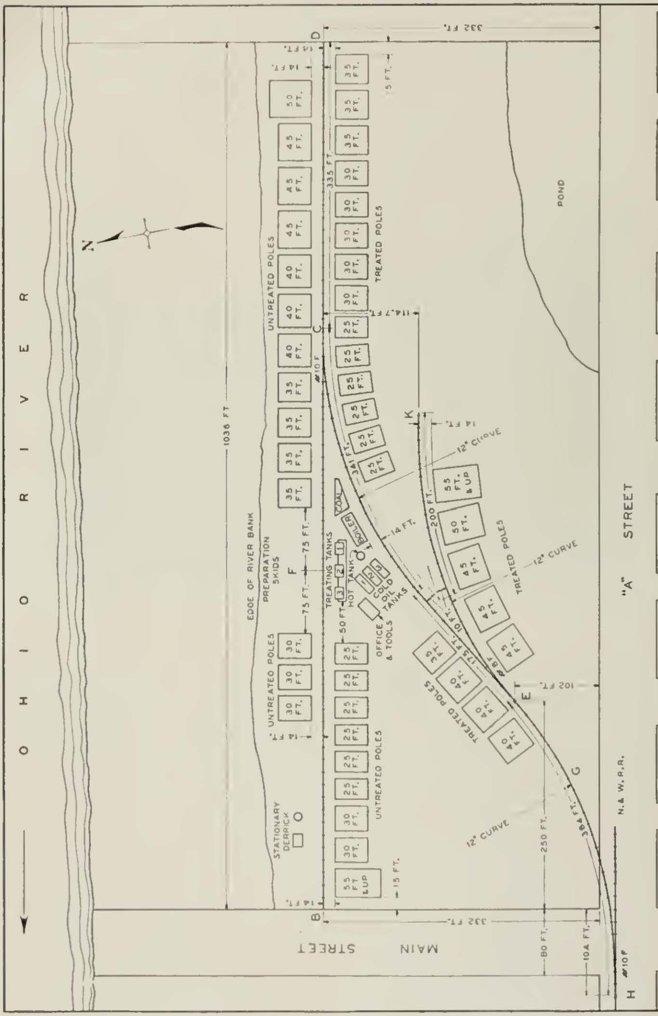


Fig. 3—Plan of Yard Layout for Ceredo. (Each Skid Shown is Separated into Two Pole Piles

Car loads of poles which are received by rail may be backed into the track leading to the pole treating plant for treatment or may be unloaded upon the "untreated" skids if desired. In any event, there should be a minimum of confusion in the pole moving operations.

Fig. 4 shows the skids at one end of the Sylva yard before poles had been piled upon them. It illustrates the desirability of having a

Fig. 4 Skid Layout at One End of Sylva Yard

long, narrow yard and also shows that the switch track is the backbone of the pole yard.

It will also be noted from Fig. 4 that in the Sylva yard the ends of the skids are brought up close to the track. This is because the pole handling in the Sylva yard is done by means of a locomotive crane which runs on the track and works from the ends of the cars.

In the Natural Bridge yard, which is shown in Fig. 5, a tractor crane is used for pole handling. This unit has crawlers and wheels which operate on the narrow roadways at either side of the spur tracks. The tractor crane runs up to the side of a car to unload it. By operating at the sides of the cars a much shorter boom is required by the tractor crane than for the locomotive crane working at the ends of the cars handling the same lengths of poles.

Delivery of Poles to Plants

Various methods are used for delivering poles to the treating plants, from the locations where they are cut. In addition to the use of automobile trucks with their trailers, and to the use of horse-drawn wagons which may be seen along the road in Fig. 4, poles are delivered by railroad cars, river rafts and ox-teams.

In the timber the poles are ordinarily loaded on cars for shipment to the treating plants by means of a logging loader shown in Fig. 6.

Fig. 5—Yard Layout at One End of Natural Bridge Yard, Viewed from Mast of Derrick

Fig. 6—Placing Poles on Logging Car by Means of Logging Loader

Although it has a short boom, it is able to handle very long poles because of the method in which it lifts them. One end of the pole, either top or butt, is rested against the middle point of the boom and the pole lifted by the winch line which may be attached only one-third or one-fourth of the distance from the loader end to the

Fig. 7—Geared Locomotive in Use on Logging Road Which Supplies Poles to Treating Plant

free end of the pole. In lifting long poles by this method, they spring considerably, and brash timber usually breaks under this treatment. Thus in handling poles by this method, they are given a test before they leave the timber.

The winch line is attached to the pole by means of hooks which resemble ice tongs. From long experience in handling these tongs, the pole men are able to throw them several feet and catch a pole at any point they desire, to pull it from the pole pile. This operation is very fast. In fact, under favorable conditions, 35 foot chestnut poles have been loaded on a car at the rate of two per minute.

The pole piles along the logging road are usually disorderly, resembling a lot of giant tooth-picks which might have been carelessly dropped in a heap.

Steep grades on the logging roads make it very desirable to use locomotives which have a maximum amount of traction. For this reason, a geared type locomotive is used which permits a big reduction between the engine and drive wheels, and also transmits the driving torque to all wheels of the engine and coal tender which is shown, and also to the wheels of the water tender which is not shown in Fig. 7.

From one to ten car loads of poles in a group arrive at the treating plants. A car load varies between 40 and 65 poles depending upon

Fig. 8—Car Load of Poles Arriving at the Danbury Treating Plant

the sizes of the poles. They may be unloaded by a locomotive crane or a tractor crane or by the method shown in Fig. 9.

At the Shipman Yard the poles are unloaded by cutting the stakes and permitting the poles to roll down an embankment into piles from which they are drawn to the treating plant by means of a steel rope from a tractor winch.

Utilization of the cheapest method of delivering poles to the treating plants is possible at Ceredo and Nashville where the plants are located on the river banks. These poles are securely tied in rafts of about 100 poles each and either floated down the rivers or handled by stern wheel, river steamboats.

Fig. 9—Unloading Poles at the Shipman Yard

Fig. 10—Four Rafts of Poles at Ceredo Plant

It may be of interest to note that the photograph shown in Fig. 10 was taken from the West Virginia bank of the river, while the Ohio bank is seen across the river and the Kentucky hills are visible beyond the bridge.

Particularly in the Carolinas, ox-teams are used to draw pole loads down from the mountains.

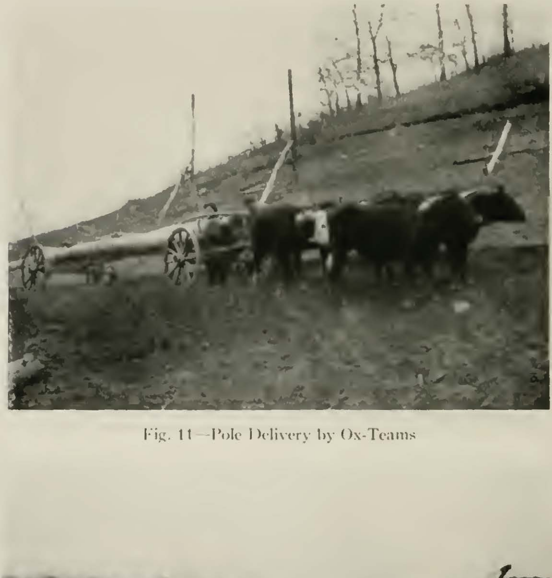


Fig. 11—Pole Delivery by Ox-Teams

Fig. 12—Derrick for Handling Poles from River Rafts to Piles or Pole Cars in the Yard

Handling Poles in the Yard

Where the derrick is used for lifting poles out of the river it is necessary to set it at a distance from the water's edge which, of course, approaches and recedes depending upon the height of the river. Because of this distance, the poles are dragged as well as lifted up the sloping side of the bank.

Fig. 13 Handling Poles by Man Power

Fig. 14 Tractor Crane Handling Poles from Rail Dollies in Danbury Yard

It has been found that wherever it is possible to eliminate the handling of poles by man-power, a considerable economy can be realized. Less men are required for crane or derrick operation, and the cranes and derricks do the work much more rapidly.

In order to move the poles about the yard it is not necessary to retain a freight car to carry them, since small rail dollies have been provided for this purpose. The two dollies shown in Fig. 14 are separate and can be located under the poles at any distance apart depending upon the lengths of the poles.

The tractor crane which is used for pole handling in the smaller plants is operated by a heavy duty gasoline engine and it is able

Fig. 15—Stiff Leg Derrick Removing Poles from Treating Tank and Loading Them on Flat Car

to handle a 1,000 lb. load at a 15 foot radius through an arc of about 270 degrees. It has a 30 foot boom. Since a very large percentage of the chestnut poles handled, weigh less than one ton each, this tractor crane has sufficient capacity for the service.

In the smaller plants where it has been found desirable to increase the pole treating capacities above what could be handled by means of the tractor cranes, stiff leg derricks have been installed. These derricks are of 6-tons capacity, having 15-foot booms. They are operated by steam from the treating plant boiler, which feeds the 8 H.P. hoisting engines. In these installations the swingers are operated by the hoisting engines.

Where the treating plant is of large enough capacity to warrant the investment in a locomotive crane, this type of unit has proven to be the most satisfactory in operation. The cranes which are suitable for this type of work have a 50 foot boom and are rated at $17\frac{1}{2}$ tons capacity. Actually they can safely handle a 3-ton load at 50 feet radius from the king pin of the crane, perpendicular to the

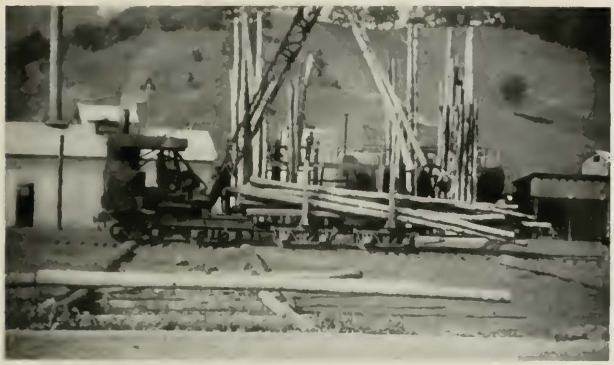


Fig. 16 Unloading Poles from the Treating Tanks to the Dollies, with Locomotive Crane

track, without tipping the car body of the crane. Of course, with the boom in a position above the track the maximum safe load is considerably greater.

The method of handling poles most commonly used is illustrated in Fig. 17 where the poles are lifted in a balanced condition, swung to one side of the track and piled parallel to it.

Another method which is applicable, particularly to handling a 40-foot and longer pole, consists of butting the pole end against the boom of the locomotive crane and swinging it to a pile which lies perpendicular to the track. This method of handling poles is similar to that shown in use with the logging outfit in Fig. 6.

When the poles are piled either parallel or perpendicular to the track as shown by Figs. 17 and 18, respectively, there should be frequent breaks in the piles in order to permit the air to circulate around the poles and keep them dry, and to reduce the fire hazard.

Fig. 17—Handling Poles by Balanced Method with Locomotive Crane

Fig. 18—Handling Pole with End Butted Against Boom of Locomotive Crane

Preparing Poles for Treatment

Although efforts were originally made to clean and prepare the poles on the cars at the time they were received at the plant, in order to be able to unload them from the cars directly into the treating tanks, it was found to be more satisfactory to first unload them upon skids where they would be more accessible for the removal

Fig. 19—Preparation Skids Opposite Treating Tanks at Sylva Plant

of all bark and foreign matter from the area to be treated and where any defective poles could be culled out before treatment.

The preparation skids are ordinarily not used for storage purposes. When a load of poles is placed upon them it can be spread in such a manner that every pole will be accessible.

In Fig. 20 the load of poles from the dollies has just been laid on the preparation skids where they will be cleaned for treatment in the far tank which is shown empty. Due to the desirability of having a continuous supply of poles for treatment, also of having the poles seasoned for several months before treatment, it is not practicable in a very large percentage of cases to ship the poles direct from the timber to the yard and unload them on the preparation skids for immediate treatment. For this reason it is necessary first to pile them in the untreated section of the pole yard and later to bring them to the preparation skids on dollies as illustrated in Fig. 20.

Treatment

The following is a very brief outline of the method pursued in treating the poles and also of the results obtained.

In so far as practicable the poles are seasoned 6 months or more before being treated. The method of treatment consists of immersing the butts to a level of about 1 foot above what will be the

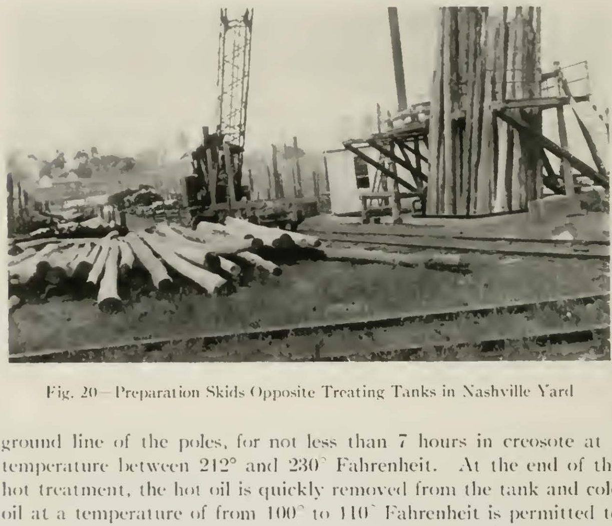


Fig. 20—Preparation Skids Opposite Treating Tanks in Nashville Yard

ground line of the poles, for not less than 7 hours in creosote at a temperature between 212° and 230° Fahrenheit. At the end of the hot treatment, the hot oil is quickly removed from the tank and cold oil at a temperature of from 100° to 110° Fahrenheit is permitted to flow quickly into the treating tank to the level previously reached by the hot oil. The cold oil treatment lasts for at least 4 hours.

Heat is absorbed by the pole butts in the hot oil bath until the moisture contained in the sapwood is either expanded into steam or entirely driven out. During the short interval while the oil is being changed, the surfaces to be treated remain covered by oil from the hot treatment. The oil change is made so quickly that the pole butts cool very little before it is completed. Then, as soon as the cold oil is admitted, these surfaces are covered by the creosote which remains until the pole butts become cool. In the sapwood, from which the moisture has been driven by the hot treatment, the cooling process condenses the steam, thus forming a partial vacuum in the

GENERAL INSTALLATION LAYOUT

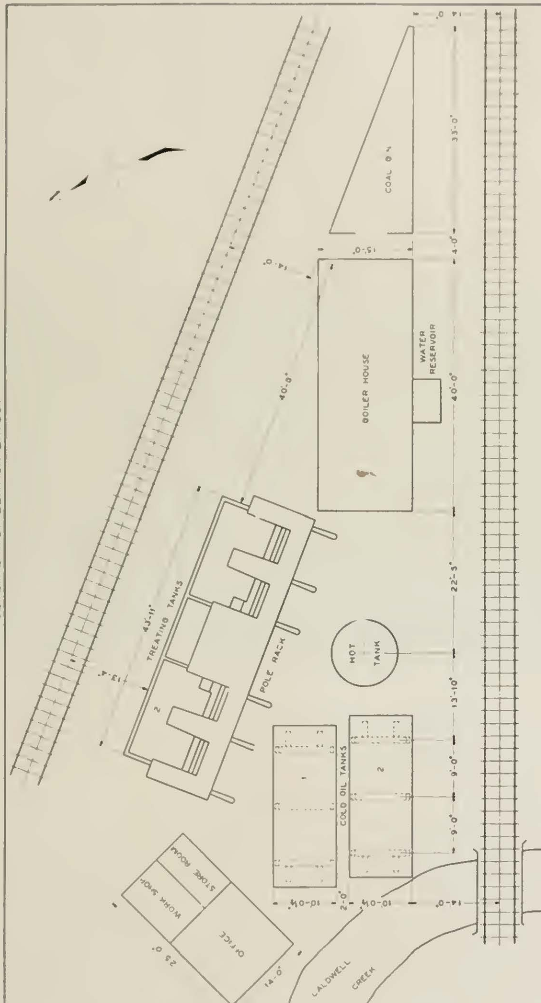


Fig. 21—Plan of Sylva Plant Layout

wood. This causes the oil, in which these surfaces are immersed, to be forced into the wood by atmospheric pressure.

During the treatment, the creosote is absorbed by the pole to such an extent that as an average, about 95 per cent. of the sapwood in the treated section of the pole is saturated. This requires from 2 to 4 gallons of oil per pole, depending upon the size and condition of the pole being treated.

Assembly Layout

The same general features of design were followed in all the pole treating plant layouts in so far as practicable. However, the number

Fig. 22—View of Treating Equipment at Sylva Plant

of the different units used was varied to provide the plant capacities required.

In designing the plants it was found desirable to separate the poles into two or three treating tanks in order that the treating gang could be continuously employed in either preparing or handling poles from or to one of the tanks while the treatment would be in progress in other tanks. By dividing the tanks it was also possible to use a smaller quantity of hot creosote, since the hot oil could be used in one tank and when that treatment was finished, pumped to another tank containing fresh poles ready for treatment. Cutting down the hot oil capacity, of course, reduced the amount of radiation in the heating tank and also the amount of radiation in use at any particular time in the treating tanks, thus resulting in considerably less steam boiler capacity than would be necessary with a very large single treating tank unit.

Handling poles at smaller tanks is much easier because less boom action of the derrick is required and the men at the tanks can reach all poles more easily for attaching and removing the derrick winch line.

It was found that a vertical cylindrical tank served better than a horizontal one for the storage of hot oil, while the horizontal cylindrical

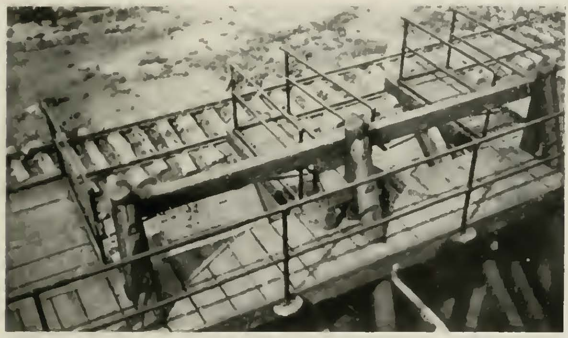


Fig. 23—Plan View of Pole Rack

tanks were preferable for cold oil storage. The radiation from a vertical hot tank is considerably reduced by the jacket of hot air rising along its side.

Particularly during the summer months care must be taken to keep down the temperature of the cold oil. It has been found that the long cylindrical steel tanks when lying horizontally radiate heat from the oil to the atmosphere satisfactorily and thus keep the oil cool.

Care has been taken in the design, to locate the various units so that all hot oil leads would be as short as possible in order to minimize radiation. Wherever possible, both the hot and cold oil are handled by gravity. The steam boiler is located as near as practicable to the heavy banks of steam radiators.

In all cases, careful study has been given to facilitating the handling of poles, since a considerable part of the cost of the pole treating process is due to pole handling.

Pole Racks

For supporting the poles standing in the treating tanks, it is necessary to have a very strong rack surrounding each of the tanks. Fig. 20 shows a view at one end and the front side of the two-tank rack in the Nashville plant. The poles shown, stand $8\frac{1}{2}$ feet below the ground level. They are supported at the ends and middle of the rack by timbers under the rack platform at a height of 12 feet above

Fig. 24 Excavation for Treating Tanks

the ground. At the back, the poles are supported by a timber which is 16 feet above the ground. This arrangement permits the treatment of any size of pole up to and including 65 feet in length.

It will be noted in Fig. 23, which shows the rack above one tank, that the poles in each tank are divided at the middle by the platform of the pole rack. This feature of the rack has proved to be very desirable in that it permits the platform man to reach any pole in the rack during the loading and unloading process, so that there is no delay and no hazard in attaching the winch line sling to, or detaching it from the poles. The taper of the poles is such that ample space is provided for holding the sections of the poles at the platform level even though the area of the opening at this level is somewhat smaller than the area of the bottom of the treating tank.

Suitable railings have been provided around all parts of the platform to protect the platform man. They are substantial enough to protect the operator and yet flexible enough to compensate for the irregular sections of poles which may lie against them.

Tanks

As was mentioned above, in so far as practicable the tanks for the various plants are made in multiples of standard units. The treating

Fig. 25—Concrete Foundation and Protecting Walls for Treating Tanks

tanks for the smaller plants are 11 feet long and 5 feet 6 inches wide with 6 inches in each end of the tanks taken up by the vertical radiators. These tanks are of proper size to treat 1/2 carload of poles each.

The larger plants are provided with treating tanks, each of which will easily handle one carload of poles. These tanks are 15 feet long, 8 feet wide and 9 feet 8 inches deep in the clear.

Some idea of the sizes and arrangement of the treating tanks can be had from the excavation for them shown in Fig. 24. Each of the raised levels shown, will support the bottom of a tank while the pits between will contain the steam and oil piping, oil handling machinery, etc. This is a three-tank pit with space for two tanks shown.

In order to provide dry pits for the equipment below the treating tank bottoms and also to facilitate removal of a tank from the ground in case it might need repair, it has been found desirable to build concrete foundations and walls around the treating tanks.

Fig. 26—Treating Tanks in Place

A few inches of space is left between the concrete retaining walls and the sides of the treating tanks. This space serves two purposes: it permits placing or removing the tanks with ease and it also provides air spaces around the sides of the tanks, which tend to insulate them from the ground. As has been mentioned, it is necessary to change the temperature of the oil in the tanks quickly from about 220° to about 105° Fahrenheit. There is very little lag in making the temperature change due to heat retained by the tank walls. However, if the ground around the tanks were wet and in contact with them, considerable lag would be experienced in making the temperature change of the oil because of heat which would be retained by the ground.

The poles in the tanks as shown by Fig. 26 rest in a position inclined slightly back toward the racks so that they remain in this position without being tied. Inclining the tank bottoms toward the rear facilitates the drainage of oil from them.

The bottom of the tank is practically perpendicular to the poles as they stand on it, which minimizes the tendency for the butts to slip on the tank bottom. In order to further prevent any danger from this happening, the bottom of each tank is covered by extra heavy Irving grids similar to those used at subway ventilating openings. These grids are supported by a suitable I-beam framework in

Fig. 27—Bottom of Treating Tank Showing Horizontal Radiators and Grids Covering Them

which the steel pipe radiators are placed. The grids do not interfere with the circulation of the hot oil and form a good protection for the radiators.

Each of the horizontal cold oil tanks has a capacity of about 14,000 gallons. Tanks of this size will easily take a tank-car load of creosote each, leaving some reserve capacity for residual oil which may be in the tanks at the time the additional cars of oil are received. The tank cars ordinarily carry from 8,000 to 12,000 gallons of oil.

The hot oil tanks vary in capacity between 3,000 and 13,000 gallons each, depending upon the sizes of the plants. One hot oil tank suffices for each installation. In order to conserve the heat, these tanks are covered by a 1½ inch coat of magnesia block heat insulating material, the outside of which is covered by ¼ inch of asbestos cement and ¼ inch of half and half asbestos and Portland cement.

Boilers, Radiators, Pressure Regulators and Other Steam Equipment

For these installations, a self-contained type of steam boiler was used because of its comparatively high efficiency in the sizes required and also because of the ease of installation. The boilers used vary

Fig. 28—Horizontal Cold Oil Tanks and Vertical Hot Oil Tank

from 30 to 80 horsepower capacity depending upon the sizes of the plants. These boilers are of the return tubular type with the fire boxes and smoke boxes lined with keyed-in fire brick.

The boilers are operated at a pressure of about 100 lbs. which is a suitable pressure for the steam turbine and for the steam hoisting engines in the plants where these are used. This boiler steam pressure is too high for the cast iron radiators which are used to heat oil in the hot and cold tanks and, for the smaller plants, in the treating tanks. Steam for these radiators should be supplied at a pressure of about 10 pounds. In order to meet this requirement a pressure reducer is used to convert the steam from the boiler pressure, whatever it may be, to a pressure of about 10 pounds, before it enters the radiators.

The water condensed from the various radiators is returned to the boiler in order to conserve its heat. Small automatic steam traps pass the water condensed in the radiators as fast as it is made, but do not permit the steam to pass. On the water side of these small steam traps, the piping from the various radiators is brought together and led to a point above the steam boiler where it is connected to a large tilting trap. The traps automatically raise the water to a

Fig. 29—Vertical Hot Oil Tank with Insulated Covering

receiver above the boiler and the tilting trap injects it into the boiler as fast as it is delivered to the water pipe lines by the small traps.

It is very desirable in the operation of the steam turbines that they be supplied with dry steam in order that slugs of water cannot enter the turbine chambers at high velocities and injure the vanes. A large water trap is located above the treating tank pit at each plant to insure dry steam for the turbine which is mounted in the pit directly below it.

Temperature Control

A continuous record is kept of the temperature of the oil in the treating tanks by means of recording thermometers mounted in the boiler room and connected by flexible thermometer tubes to the bulbs which are immersed in oil along the inside of the tanks after the poles are in place. In the cold and hot tanks the temperature does not change rapidly, so their temperatures can be read by means of stationary indicating thermometers mounted on the sides of the tanks and having bulbs which project into the insides of the tanks through suitable fittings. The oil temperatures, of course, are controlled by the steam valves to the radiators in the various tanks.

Fig. 30—Steam Boiler During Installation

Oil Handling

The heart of the oil handling apparatus, of course, is the centrifugal pump which has been mentioned and which is direct connected to the 20 H.P. steam turbine. In some of the smaller plants the centrifugal pumps are operated by 5 H.P. gasoline engines.

Both cold oil and hot oil are fed from the storage tanks to the treating tanks by gravity. The centrifugal pump is used for returning the oil from the treating tanks to the proper storage tank, for moving it from one storage tank to another or for delivering oil from the tank cars to the storage tanks.

Since the creosote which is used in pole treating may solidify at any temperature below 100° Fahrenheit, even in comparatively warm weather it is sometimes necessary to provide a steam connection to the radiators inside the tank car in order to make the oil fluid enough to flow through the flexible hose and pipes to the centrifugal pump. The solidifying of the creosote at comparatively high temperatures also requires a small bank of radiators in each cold tank.

The steam pipe runs, between the steam boiler and the various tanks, and the oil pipe lines between the various tanks and the pump,

Fig. 31 General View of Natural Bridge Plant in Operation

are grouped so that both the steam lines and oil lines can be enclosed in boxes. The heat radiated from the steam lines warms the air in the boxes to such an extent that the oil remains liquid.

The valve controls for the oil and steam lines which are led through the boxes, are grouped so that several can be reached by opening the door of each of the boxes.

In the smaller plants which have the one-half-car pole capacity of treating tanks, the centrifugal pump handles the oil at a rate of about 200 gallons per minute. In the larger plants, however, where the treating tanks have one-car capacity of poles, the oil is handled through the centrifugal pump at the rate of about 600 gallons per minute. As mentioned in the above section describing the treatment, the high rate of oil movement is necessary in order to accomplish the change from hot to cold oil in the treating tanks in such a short time that the heated pole butts will not be permitted to cool when not immersed in oil. The oil change ordinarily is made in from 7 to 12 minutes from the time the pump starts to remove the hot oil until the cold oil is up to the proper level.

Experience indicates that no material loss in penetration of the creosote into the poles is experienced by having the treated section uncovered for this short length of time. Practically the same penetration is obtained as would be secured by keeping the poles in hot oil for the same length of time and then permitting the hot oil to remain around them until its temperature had gradually fallen by radiation to that specified for the cold oil bath.

Changing the oil instead of permitting it to cool in the treating tanks greatly expedites the treatments and consequently increases the plant capacity, which, of course, results in a corresponding economy in the cost of treating the poles.

Conclusion

In this paper an endeavor has been made to cover in a general way, the principal engineering and operating features involved in building creosoting plants designed specially for applying open tank treatment to chestnut poles. It has, of course, been necessary to omit practically all of the details of construction, which were followed in building the various plants.

These treating plants are valuable assets to the Bell System in providing concentration points where preservative treatment can be economically applied to the chestnut poles, thus becoming an important factor in the general program for the conservation of natural resources, by making possible the utilization of this valuable and rapidly diminishing type of timber over a considerably longer period.

Selective Circuits and Static Interference*

By JOHN R. CARSON

SYNOPSIS: The present paper has its inception in the need of a correct understanding of the behavior of selective circuits when subjected to irregular and random interference, and of devising a practically useful figure of merit for comparing circuits designed to reduce the effects of this type of interference. The problem is essentially a statistical one and the results must be expressed in terms of mean values. The mathematical theory is developed from the idea of the spectrum of the interference and the response of the selective circuit is expressed in terms of the mean square current and mean power absorbed. The application of the formulas deduced to the case of static interference is discussed and it is shown that deductions of practical value are possible in spite of meagre information regarding the precise nature and origin of static interference.

The outstanding deductions of practical value may be summarized as follows:

1. Even with absolutely ideal selective circuits, an irreducible minimum of interference will be absorbed, and this minimum increases linearly with the frequency range necessary for signaling.

2. The wave-filter, when properly designed, approximates quite closely to the ideal selective circuit, and little, if any, improvement over its present form may be expected as regards static interference.

3. As regards static or random interference, it is quite useless to employ extremely high selectivity. The gain, as compared with circuits of only moderate selectivity, is very small, and is inevitably accompanied by disadvantages such as sluggishness of response with consequent slowing down of the possible speed of signaling.

4. A formula is developed, which, together with relatively simple experimental data, provides for the accurate determination of the spectrum of static interference.

5. An application of the theory and formulas of the paper to representative circuit arrangements and schemes designed to reduce static interference, shows that they are incapable of reducing, in any substantial degree, the mean interference, as compared with what can be done with simple filters and tuned circuits. The underlying reason lies in the nature of the interference itself.

I

THE selective circuit is an extremely important element of every radio receiving set, and on its efficient design and operation depends the economical use of the available frequency range. The theory and design of selective circuits, particularly of their most conspicuous and important type, the electric wave filter, have been highly developed, and it is now possible to communicate simultaneously without undue interference on neighboring channels with a quite small frequency separation. On the other hand too much has been expected of the selective circuit in the way of eliminating types of interference which inherently do not admit of elimination by any form of selective circuit. I refer to the large amount of inventive thought devoted to devising ingenious and complicated circuit ar-

* Presented at the Annual Convention of the A. I. E. E., Edgewater Beach, Chicago, Ill., June 23-27, 1924.

rangements designed to eliminate *static interference.* Work on this problem has been for the most part futile, on account of the lack of a clear analysis of the problem and a failure to perceive inherent limitations on its solutions by means of selective circuits.

The object of this paper is twofold: (1) To develop the mathematical theory of the behavior of selective circuits when subjected to random, irregular disturbances, hereinafter defined and designated as *random interference.* This will include a formula which is proposed as a measure of the *figure of merit of selective circuits with respect to random interference.* (2) On the basis of this theory to examine the problem of *static interference* with particular reference to the question of its elimination by means of selective circuits. The mathematical theory shows, as might be expected, that the complete solution of this problem requires experimental data regarding the frequency distribution of static interference which is now lacking. On the other hand, it throws a great deal of light on the whole problem and supplies a formula which furnishes the theoretical basis for an actual determination of the spectrum of static. Furthermore, on the basis of a certain mild and physically reasonable assumption, it makes possible general deductions of practical value which are certainly qualitatively correct and are believed to involve no quantitatively serious error. These conclusions, it may be stated, are in general agreement with the large, though unsystematized, body of information regarding the behavior of selective circuits to static interference, and with the meagre data available regarding the wave form of elementary static disturbances.

The outstanding conclusions of practical value of the present study may be summarized as follows:

(1) Even with absolutely ideal selective circuits, an irreducible minimum of interference will be absorbed, and this minimum increases linearly with the frequency range necessary for signaling.

(2) The wave-filter, when properly designed, approximates quite closely to the ideal selective circuit, and little, if any, improvement over its present form may be expected as regards static interference.

(3) As regards static or random interference, it is quite useless to employ extremely high selectivity. The gain, as compared with circuits of only moderate selectivity, is very small, and is inevitably accompanied by disadvantages such as sluggishness of response with consequent slowing down of the possible speed of signaling.

(4) By aid of a simple, easily computed formula, it should be possible to determine experimentally the frequency spectrum of static.

(5) Formulas given below for comparing the relative efficiencies of selective circuits on the basis of signal-to-interference energy ratio are believed to have considerable practical value in estimating the relative utility of selective circuits as regards static interference.

II

Discrimination between signal and interference by means of selective circuits depends on taking advantage of differences in their wave forms, and hence on differences in their *frequency spectra*. It is therefore the function of the selective circuit to respond effectively to the range of frequencies essential to the signal while discriminating against all other frequencies.

Interference in radio and wire communication may be broadly classified as *systematic* and *random*, although no absolutely hard and fast distinctions are possible. *Systematic interference* includes those disturbances which are predominantly steady-state or those whose energy is almost all contained in a relatively narrow band of the frequency range. For example, interference from individual radiotelephone and slow-speed radio telegraph stations is to be classified as systematic. *Random interference*, which is discussed in detail later, may be provisionally defined as the aggregate of a large number of elementary disturbances which originate in a large number of unrelated sources, vary in an irregular, arbitrary manner, and are characterized statistically by no sharply predominate frequency. An intermediate type of interference, which may be termed either *quasi-systematic* or *quasi-random*, depending on the point of view, is the aggregate of a large number of individual disturbances, all of the same wave form, but having an irregular or random time distribution.

In the present paper we shall be largely concerned with random interference, as defined above, because it is believed that it represents more or less closely the general character of *static* interference. This question may be left for the present, however, with the remark that the subsequent analysis shows that, as regards important practical applications and deductions, a knowledge of the exact nature and frequency distribution of static interference is not necessary.

Now when dealing with random disturbance, as defined above, no information whatsoever is furnished as regards instantaneous values. In its essence, therefore, the problem is a statistical one and the conclusions must be expressed in terms of mean values. In the present paper formulas will be derived for the *mean energy* and *mean square current* absorbed by selective circuits from random interference, and their applications to the static problem and the protection afforded by selective networks against static will be discussed.

The analysis takes its start with certain general formulas given by the writer in a recent paper[1], which may be stated as follows:

Suppose that a selective network is subjected to an impressed force $\phi\ (t)$. We shall suppose that this force exists only in the time interval, or epoch, $o \leqq t \leqq T$, during which it is everywhere finite and has only a finite number of discontinuities and a finite number of maxima and minima. It is then representable by the Fourier Integral

$$\phi(t) = 1/\pi \int_0^\infty |f(\omega)| \cdot \cos\left[\omega t + \theta(\omega)\right] d\omega \qquad (1)$$

where

$$|f(\omega)|^2 = \left[\int_0^\infty \phi(t)\ \cos\ \omega t dt\right]^2 + \left[\int_0^\infty \phi(t)\ \sin\ \omega t dt\right]^2. \qquad (2)$$

Now let this force $\phi\ (t)$ be applied to the network in the *driving* branch and let the resulting current in the *receiving* branch be denoted by $I\ (t)$. Let $Z\ (i\ \omega)$ denote the steady-state *transfer* impedance of the network at frequency $\omega/2\pi$: that is the ratio of e.m.f. in *driving* branch to current in *receiving* branch. Further let $z\ (i\ \omega)$ and $\cos\ \alpha\ (\omega)$ denote the corresponding impedance and power factor of the receiving branch. It may then be shown that

$$\int_0^\infty [I(t)]^2 dt = 1/\pi \int_0^\infty \frac{|f\ (\omega)|^2}{|Z(i\omega)|^2} d\omega \qquad (3)$$

and that the total energy W absorbed by the receiving branch is given by

$$W = 1/\pi \int_0^\infty \frac{|f(\omega)|^2}{|Z(i\omega)|^2} |z(i\omega)|\ \cos\alpha(\omega) \cdot d\omega. \qquad (4)$$

To apply the formulas given above to the problem of random interference, consider a time interval, or epoch, say from $t = o$ to $t = T$, during which the network is subjected to a disturbance made up of a large number of unrelated elementary disturbances or forces, $\phi_1\ (t)$, $\phi_2\ (t)\ .\ .\ .\ \phi_n\ (t)$.

If we write

$$\Phi(t) = \phi_1(t) + \phi_2(t) + \ .\ .\ .\ + \phi_n(t),$$

then by (1), $\Phi(t)$ can be represented as

$$\Phi(t) = 1/\pi \int_0^\infty |F(\omega)| \cdot \cos\left[\omega t + \theta(\omega)\right] d\omega$$

[1] Transient Oscillations in Electric Wave Filters, Carson and Zobel, *Bell System Technical Journal*, July, 1923.

and

$$\int_0^\infty [I(t)]^2 dt = 1/\pi \int_0^\infty \frac{|F(\omega)|^2}{|Z(i\omega)|^2} d\omega. \tag{3}$$

We now introduce the function $R(\omega)$, which will be termed the *energy spectrum* of the random interference, and which is analytically defined by the equation

$$R(\omega) = \frac{1}{T} |F(\omega)|^2 \tag{5}$$

Dividing both sides of (3) and (4) by T we get

$$\overline{I^2} = 1/\pi \int_0^\infty \frac{R(\omega)}{|Z(i\omega)|^2} d\omega, \tag{6}$$

$$\overline{P} = 1/\pi \int_0^\infty \frac{R(\omega)}{|Z(i\omega)|^2} |z(i\omega)| \cdot \cos \alpha(\omega) \, . \, d\omega. \tag{7}$$

$\overline{I^2}$, $\overline{P}$ and $R(\omega)$ become independent of the T provided the epoch is made sufficiently great. *$\overline{I^2}$ is the mean square current and $\overline{P}$ the mean power absorbed by the receiving branch from the random interference.*

In the applications of the foregoing formulas to the problem under discussion, the mean square current $\overline{I^2}$ of the formula (6) will be taken as the relative measure of interference instead of the mean power $\overline{P}$ of formula (7). The reason for this is the superior simplicity, both as regards interpretation and computation, of formula (6). The adoption of $\overline{I^2}$ as the criterion of interference may be justified as follows:

(1) In a great many important cases, including in particular experimental arrangements for the measurement of the static energy spectrum, the receiving device is substantially a pure resistance. In such cases multiplication of $\overline{I^2}$ by a constant gives the actual mean power $\overline{P}$.

(2) It is often convenient and desirable in comparing selective networks to have a standard termination and receiving device. A three-element vacuum tube with a pure resistance output impedance suggests itself, and for this arrangement formulas (6) and (7) are equal within a constant.

(3) We are usually concerned with relative amounts of energy absorbed from static as compared with that absorbed from signal. Variation of the receiver impedance from a pure constant resistance would only in the extreme cases affect this ratio to any great extent. In other words, the ratio calculated from formula (6) would not differ greatly from the ratio calculated from (7).

(4) While the interference actually apperceived either visually or by ear will certainly depend upon and increase with the energy absorbed from static, it is not at all certain that it increases linearly therewith. Consequently, it is believed that the additional refinement of formula (7) as compared with formula (6) is not justified by our present knowledge and that the representation of the receiving device as a pure constant resistance is sufficiently accurate for present purposes. It will be understood, however, that throughout the following argument and formulas, P of formula (7) may be substituted for $\bar{I}^2$ of (6), when the additional refinement seems justified. The theory is in no sense limited to the idea of a pure constant resistance receiver, although the simplicity of the formulas and their ease of computation is considerably enhanced thereby.

The problem of random interference, as formulated by equations (6) and (7) was briefly discussed by the writer in "Transient Oscillations in Electric Wave Filters" [1] and a number of general conclusions arrived at. That discussion will be briefly summarized, after which a more detailed analysis of the problem will be given.

Referring to formula (6), since both numerator and denominator of the integrand are everywhere $\geqq o$, it follows from the mean value theorem that a value ω of ω exists such that

$$\bar{I}^2 = \frac{R(\omega)}{\pi}\int_0^\infty \frac{d\omega}{|Z(i\omega)|^2}. \tag{8}$$

The approximate location of ω on the frequency scale is based on the following considerations:

(a) In the case of efficient selective circuits designed to select a continuous finite range of frequencies in the interval $\omega_1 \leqq \omega \leqq \omega_2$, the important contributions to the integral (6) are confined to a finite continuous range of frequencies which includes, but is not greatly in excess of, the range which the circuit is designed to select. This fact is a consequence of the impedance characteristics of selective circuits, and the following properties of the spectrum $R(\omega)$ of random interference, which are discussed in detail subsequently.

(b) $R(\omega)$ is a continuous finite function of ω which converges to zero at infinity and is everywhere positive. It possesses no sharp maxima or minima, and its variation with respect to ω, where it exists, is relatively slow.

On the basis of these considerations it will be assumed that ω lies within the band $\omega_1 \leqq \omega \leqq \omega_2$ and that without serious error it may be taken as the mid-frequency ω_m of the band which may be defined either as $(\omega_1+\omega_2)/2$ or as $\sqrt{\omega_1\omega_2}$. Consequently

$$\bar{I}^2=\frac{R(\omega_m)}{\pi}\int_0^\infty \frac{d\omega}{|Z(i\omega)|^2}. \tag{9}$$

From (9) it follows that the mean square current $\bar{I}^2$, due to random interference, is made up of two factors: one $R\ (\omega_m)$ which is proportional to the energy level of the interference spectrum at mid-frequency $\omega_m/2\pi$: and, second, the integral

$$\rho=1/\pi\int_0^\infty \frac{d\omega}{|Z\ (i\omega)|^2} \tag{10}$$

which is independent of the character and intensity of the interference. Thus

$$\bar{I}^2=\rho R(\omega_m). \tag{11}$$

Formula (11) is of considerable practical importance, because by its aid the spectral energy level $R\ (\omega)$ can be determined, once $\bar{I}^2$ is experimentally measured and the frequency characteristics of the receiving network specified or measured. It is approximate, as discussed above, but can be made as accurate as desired by employing a sufficiently sharply selective network.

The formula for the *figure of merit of a selective circuit with respect to random interference* is constructed as follows:

Let the signaling energy be supposed to be spread continuously and uniformly over the frequency interval corresponding to $\omega_1 \leqq \omega \leqq \omega_2$. Then the mean square signal current is given by

$$\frac{E^2}{\pi}\int_{\omega_1}^{\omega_2} \frac{d\ \omega}{|Z(i\omega)|^2}$$

or, rather, on the basis of the same transmitted energy to

$$\frac{E^2}{\pi(\omega_2-\omega_1)}\int_{\omega_1}^{\omega_2} \frac{d\omega}{|Z(i\omega)|^2}=E^2\frac{\sigma}{\omega_2-\omega_1}. \tag{12}$$

The ratio of the mean square currents, due to signal and to interference, is

$$\frac{E^2}{R(\omega_m)}\cdot\frac{1}{\omega_2-\omega_1}\frac{\sigma}{\rho}. \tag{13}$$

The first factor $\frac{E^2}{R(\omega_m)}$ depends only on the signal and interference energy levels, and does not involve the properties of the network. The second factor depends only on the network and measures the efficiency with which it excludes energy outside the signaling range. It will therefore be termed *the figure of merit of the selective circuit* and denoted by S, thus

$$S = \frac{1}{\omega_2 - \omega_1} \frac{\sigma}{\rho} = \frac{1}{\omega_2 - \omega_1} \int_{\omega_1}^{\omega_2} \frac{d\omega}{|Z(i\omega)|^2} \div \int_0^\infty \frac{d\omega}{|Z(i\omega)|^2}. \qquad (14)$$

Stated in words, *the figure of merit of a selective circuit with respect to random interference is equal to the ratio of the mean square signal and interference currents in the receiver, divided by the corresponding ratio in an ideal band filter which transmits without loss all currents in a "unit" band* ($\omega_2 - \omega_1 = 1$) *and absolutely extinguishes currents outside this band.*

III

Before taking up practical applications of the foregoing formulas further consideration will be given to the hypothesis, fundamental to the argument, that over the frequency range which includes the important contributions to the integral $\int_0^\infty \frac{d\omega}{|Z(i\omega)|^2}$ the spectrum $R(\omega)$ has negligible fluctuations so that the integral

$$\int_0^\infty \frac{R(\omega)}{|Z(i\omega)|^2} d\omega$$

may, without appreciable error, be replaced by

$$R(\omega_m) \int_0^\infty \frac{d\omega}{|Z(i\omega)|^2}$$

where $\omega_m/2\pi$ is the "mid-frequency" of the selective circuit.

The original argument in support of this hypothesis was to the effect that, since the interference is made up of a large number of unrelated elementary disturbances distributed at random in time, any sharp maxima or minima in the spectrum of the individual disturbances would be smoothed out in the spectrum of the aggregate disturbance. This argument is still believed to be quite sound: the importance of the question, however, certainly calls for the more detailed analysis which follows:

Let
$$\Phi(t) = \sum_1^N \phi_r(t - t_r) \qquad (15)$$

where t_r denotes the time of incidence of the r^{th} disturbance $\phi_r(t)$. The elementary disturbances $\phi_1, \phi_2 \ldots \phi_N$ are all perfectly arbitrary, so that $\Phi(t)$ as defined by (15) is the most general type of disturbance possible. The only assumption made as yet is that the instants of incidence $t_1 \ldots t_N$ are distributed at random over the epoch $o \leqq t \leqq T$; an assumption which is clearly in accordance with the facts in the case of static interference. If we write

$$C_r(\omega) = \int_0^\infty \phi_r(t) \cos \omega t \, dt,$$

$$S_r(\omega) = \int_0^\infty \phi_r(t) \sin \omega t \, dt, \tag{16}$$

it follows from (2) and (15), after some easy rearrangements that

$$|F(\omega)|^2 = \sum_{r=1}^{N} \sum_{s=1}^{N} \cos \omega(t_r - t_s)[C_r(\omega) C_s(\omega) + S_r(\omega) S_s(\omega)] =$$

$$\sum C_r^2(\omega) + S_r^2(\omega) \tag{17}$$

$$+ \sum \sum \cos \omega(t_r - t_s) \, [C_r(\omega) C_s(\omega) + S_r(\omega) \, S_s(\omega)], \; r \neq s.$$

The first summation is simply $\sum |f_r(\omega)|^2$. The double summation involves the factor $\cos \omega \, (t_r - t_s)$. Now by virtue of the assumption of random time distribution of the elementary disturbances, it follows that t_r and t_s, which are independent, may each lie anywhere in the epoch $o \leqq t \leqq T$ with all values equally likely. The mean value of $|F(\omega)|^2$ is therefore gotten by averaging[2] with respect to t_r and t_s over all possible values, whence

$$\overline{|F(\omega)|^2} = \sum |f_r(\omega)|^2 + \frac{2}{T^2} \frac{1 - \cos \omega T}{\omega^2}$$

$$\times \sum \sum [C_r(\omega) C_s(\omega) + S_r(\omega) S_s(\omega)] \tag{18}$$

and

$$\overline{I^2} = \frac{1}{\pi T} \sum \int_0^\infty \frac{|f_r(\omega)|^2}{|Z(i\omega)|^2} d\omega + \frac{2}{\pi T^2} \sum \sum \int_0^\infty \frac{1 - \cos \omega T}{\omega^2 T} [C_r(\omega) C_s(\omega)$$

$$+ S_r(\omega) S_s(\omega)] \, \frac{d\omega}{|Z(i\omega)|^2}.$$

[2] The averaging process with respect to the parameters t_r and t_s employed above logically applies to the average result in a very large number of epochs during which the system is exposed to the same set of disturbances with different but random time distributions. Otherwise stated, the averaging process gives the mean value corresponding to all possible equally likely times of incidence of the elementary disturbances. The assumption is, therefore, that if the epoch is made sufficiently large, the actual effect of the unrelated elementary disturbances will in the long run be the same as the average effect of all possible and equally likely distributions of the elementary disturbances.

Now in the double summation if the epoch T is made sufficiently great, the factor $\frac{(1-\cos\omega T)}{\omega^2 T}$ vanishes everywhere except in the neighborhood of $\omega = o$. Consequently, the double summation can be written as

$$\frac{2}{\pi T^2}\int_0^\infty \frac{1-\cos\omega T}{\omega^2 T^2}\, d\omega T \cdot \sum\sum \frac{C_r(o)C_s(o)}{|Z(o)|^2} = \frac{1}{T^2}\sum\sum\frac{C_r(o)C_s(o)}{|Z(o)|^2}.$$

Finally if we write $N/T = n =$ average number of disturbances per unit time, and make use of formula (2), we get

$$\bar{I}^2 = \frac{n}{N}\sum 1/\pi\int_0^\infty \frac{|f_r(\omega)|^2}{|Z(i\omega)|^2}d\omega + \frac{n^2}{N^2}\cdot\frac{1}{|Z(o)|^2}\cdot\sum\sum\int_0^\infty \phi_r(t)dt\cdot\int_0^\infty\phi_s(t)dt, \tag{19}$$

which can also be written as

$$\bar{I}^2 = \frac{n}{N}\sum\int_0^\infty i_r^2 dt + \frac{n^2}{N^2}\sum\sum\int_0^\infty i_r dt\cdot\int_0^\infty i_s dt\cdot \tag{20}$$

when $i_r = i_r\,(t)$ is the current due to the r^{th} disturbance $\phi_r\,(t)$.

Now the double summation vanishes when, due to the presence of a condense or transformer, the circuit does not transmit direct current to the receiving branch. Furthermore, if the disturbances are oscillatory or alternate in sign at random, it will be negligibly small compared with the single summation. Consequently, it is of negligible significance in the practical applications contemplated, and will be omitted except in special cases. Therefore, disregarding the double summation, the foregoing analysis may be summarized as follows:

$$R(\omega) = \frac{n}{N}\sum|f_r(\omega)|^2 = n\cdot r(\omega), \tag{21}$$

$$\bar{I}^2 = \frac{n}{N}\sum 1/\pi\int_0^\infty \frac{|f_r(\omega)|^2}{|Z(i\omega)|^2}d\omega \tag{22}$$

$$= \frac{n}{N}\sum\int_0^\infty i_r^2 dt = n\int_0^\infty \bar{i}^2 dt, \tag{23}$$

$$P = \frac{n}{N}\int_0^\infty \frac{r(\omega)}{|Z(i\omega)|^2}|z(i\omega)|\cdot\cos\alpha\,(\omega)\cdot d\omega \tag{24}$$

$$= \frac{n}{N}\sum w_r = n\cdot w. \tag{25}$$

In these formulas n denotes the average number of elementary disturbances per unit time, w_m the energy absorbed from the r^{th} disturbance $\phi_r(t)$, and P the mean power absorbed from the aggregate disturbance. $r(\omega)$ is defined by formula (20) and is the mean spectrum of the aggregate disturbance, thus

$$r(\omega) = 1/N \sum |f_r(\omega)|^2 = R(\omega)/N. \tag{26}$$

We are now in a position to discuss more precisely the approximations, fundamental to formulas (9)–(11),

$$\int_0^\infty \frac{R(\omega)}{|Z(i\omega)|^2} d\omega = R(\omega_m) \int_0^\infty \frac{d\omega}{|Z(i\omega)|^2}. \tag{27}$$

The approximation involved in this formula consists in identifying $\omega_m/2\pi$ with the "mid-frequency" of the selective circuit, and is based on the hypothesis that over the range of frequencies, which includes the important contribution to the integral (22), the fluctuation of $R(\omega)$ may be ignored.

Now it is evident from formulas (21)–(22) that the theoretically complete solution of the problem requires that $R(\omega)$ be specified over the entire frequency range from $\omega = o$ to $\omega = \infty$. Obviously, the required information cannot be deduced without making some additional hypothesis regarding the character of the interference or the mechanism in which it originates. On the other hand, the mere assumption that the individual elementary disturbances $\phi_1 \ldots \phi_N$ differ among themselves substantially in wave form and duration, or that the maxima of the corresponding spectra $|f_r(\omega)|$ are distributed over a considerable frequency range, is sufficient to establish the conclusion that the individual fluctuations are smoothed out in the aggregate and that consequently $r(\omega)$ and hence $R(\omega)$ would have negligible fluctuations, or curvature with respect to ω, over any limited range of frequencies comparable to a signaling range.

It is admitted, of course, that the foregoing statements are purely qualitative, as they must be in the absence of any precise information regarding the wave forms of the elementary disturbances constituting random interference. On the other hand, the fact that static is encountered at all frequencies without any sharp changes in its intensity as the frequency is varied, and that the assumption of a systematic wave form for the elementary disturbances would be physically unreasonable, constitute strong inferential support of the hypothesis underlying equation (27). Watt and Appleton (*Proc. Roy. Soc.*, April 3, 1923) supply the only experimental data regarding the wave forms of the elementary disturbances which they found to be classifiable under general types with rather widely variable amplitudes and durations. Rough calculations of $r(\omega)$, based on their results, are in support of the hypothesis made in this paper, at least in the radio frequency range. In addition, the writer has made calculations based on a number of reasonable assumptions regarding variations of wave form among the individual disturbances, all of which resulted in a spectrum $R(\omega)$ of negligible fluctuations over a frequency range necessary to justify equation (27) for efficient selective circuits. However the problem is not theoretically solvable by pure mathematical analysis, so that the rigorous verification of the theory of selectivity developed in this paper must be based on experimental evidence. On the other hand, it is submitted that the hypothesis introduced regarding static interference is not such as to vitiate the conclusions, qualitatively considered, or in general to introduce serious quantitative errors. Furthermore, even if it were admitted for the sake of argument that the figure of merit S was not an accurate measure of the ratio of mean square signal to interference current, nevertheless, it is a true measure of the excellence of the circuit in excluding interference energy outside the necessary frequency range.

IV

The practical applications of the foregoing analysis depend upon the formulas

$$\bar{I}^2 = \frac{R(\omega_m)}{\pi}\int_0^\infty \frac{d\omega}{|Z(i\omega)|^2} = \rho \cdot R(\omega_m) \tag{11}$$

and

$$S = \frac{1}{\omega_2 - \omega_1} \cdot \int_{\omega_1}^{\omega_2} \frac{d\omega}{|Z(i\omega)|^2} \div \int_0^\infty \frac{d\omega}{|Z(i\omega)|^2} = \frac{1}{\omega_2 - \omega_1} \cdot \frac{\sigma}{\rho} \tag{14}$$

which contain all the information which it is possible to deduce in the case of purely random interference. They are based on the principle that the effect of the interference on the signaling system is measured by the mean square interference current in the receiving branch, and that the efficiency of the selective circuit is measured by the ratio of the mean square signal and interference currents. As stated above, in the case of random interference results must be expressed in terms of mean values, and it is clear that either the mean square current or the mean energy is a fundamental and logical criterion.

Referring to formula (11), the following important proposition is deducible.

If the signaling system requires the transmissions of a band of frequencies corresponding to the interval $\omega_2 - \omega_1$, *and if the selective circuit is efficiently designed to this end, then the mean square interference current is proportional to the frequency band width* $\frac{(\omega_2 - \omega_1)}{2\pi}$.

This follows from the fact that, in the case of efficiently designed band-filters, designed to select the frequency range $\frac{(\omega_2 - \omega_1)}{2\pi}$ and exclude other frequencies, the integral $\int_0^\infty \frac{d\omega}{|Z(i\omega)|^2}$ is proportional to $\omega_2 - \omega_1$ to a high degree of approximation.

The practical consequences of these propositions are important and immediate. It follows that as the signaling speed is increased, the amount of interference inevitably increases practically linearly and that this increase is inherent. Again it shows the advantage of single vs. double side-band transmission in carrier telephony, as pointed out by the writer in a recent paper.[3] It should be noted that the increased interference with increased signaling band width is not due to any failure of the selective circuit to exclude energy outside the signaling range, but to the inherent necessity of absorbing the interference energy lying inside this range. The only way in which the interference can be reduced, assuming an efficiently designed band filter and a prescribed frequency range $\frac{(\omega_2 - \omega_1)}{2\pi}$, is to select a carrier frequency, at which the energy spectrum $R(\omega)$ of the interference is low.

Formula (11) *provides the theoretical basis for an actual determination of the static spectrum.* Measurement of I^2 over a sufficiently long interval, together with the measured or calculated data for evaluating the integral $\int_0^\infty \frac{d\omega}{|Z(i\omega)|^2}$, determines $R(\omega_m)$ and this determination can be made as accurate as desired by employing a sufficiently sharply tuned circuit or a sufficiently narrow band filter. It is suggested that the experimental data could be gotten without great difficulty, and that the resulting information regarding the statistical frequency distribution of static would be of large practical value.

The selective figure of merit S as defined by (14) is made up of two factors, $\frac{1}{(\omega_2 - \omega_1)}$ which is inversely proportional to the required signaling frequency range; and the ratio of the integrals σ/ρ. This

[3] Signal-to-Static-Interference Ratio in Radio Telephony, *Proc. I. R. E. E.*, June, 1923.

ratio is unity for an ideally designed selective circuit, and can actually be made to approximate closely to unity with correctly designed band-filters. Formula (14) is believed to have very considerable value in comparing various circuits designed to eliminate interference, and is easily computed graphically when the frequency characteristics of the selective circuit are specified.

The general propositions deducible from it may be briefly listed and discussed as follows:

With a signaling frequency range $\frac{(\omega_2-\omega_1)}{2\pi}$ *specified, the upper limiting value of S with a theoretically ideal selective circuit is* $\frac{1}{(\omega_2-\omega_1)}$, *and the excellence of the actual circuit is measured by the closeness with which its figure of merit approaches this limiting value.*

Formula (14) for the figure of merit S has been applied to the study of the optimum design of selective circuits and to an analysis of a large number of arrangements designed to eliminate or reduce static interference. The outstanding conclusions from this study may be briefly reviewed and summarized as follows:

The form of the integrals σ and ρ, taking into account the signaling requirements, shows that the optimum selective circuit, as measured by S, is one which has a constant transfer impedance over the signaling frequency range $\frac{(\omega_2-\omega_1)}{2\pi}$, and attenuates as sharply as possible currents of all frequencies outside this range. Now this is precisely the ideal to which the band filter, when properly designed and terminated, closely approximates, and leads to the inference that *the wave filter is the best possible form of selective circuit, as regards random interference.* Its superiority from the steady-state viewpoint has, of course, long been known.

An investigation of the effect of securing extremely high selectivity by means of filters of a large number of sections was made, and led to the following conclusion:

In the case of an efficiently designed band-filter, terminated in the proper resistance to substantially eliminate reflection losses, the figure of merit is given to a good approximation by the equation

$$S=\frac{1}{\omega_2-\omega_1}\,\frac{1}{1+1/16n^2}$$

where n is the number of filter sections and $\frac{(\omega_2-\omega_1)}{2\pi}$ the transmission band. It follows that *the selective figure of merit increases inappreciably with an increase in the number of filter sections beyond 2, and that the* *band filter of a few sections can be designed to have a figure of merit closely approximating the ideal limiting value,* $\frac{1}{(\omega_2-\omega_1)}$.

This proposition is merely a special case of the general principle that, as regards static interference, it is useless to employ extremely high selectivity. The gain obtainable, as compared with only a moderate amount of selectivity is slight and is inherently accompanied by an increased sluggishness of the circuit. That is to say, as the selectivity is increased, the time required for the signals to build up is increased, with a reduction in quality and possible signaling speed.

Another circuit of practical interest, which has been proposed as a solution of the "static" problem in radio-communication consists of a series of sharply tuned oscillation circuits, unilaterally coupled through amplifiers.[4] This circuit is designed to receive only a single frequency to which all the individual oscillation circuits are tuned. The figure of merit of this circuit is approximately

$$S = \frac{L}{R}\,\frac{2^{2n-2}\,(n-1)!^2}{(2n-2)!}$$

where n denotes the number of sections or stages, and L and R are the inductance and resistance of the individual oscillation circuits. The outstanding fact in this formula is the slow rate of increase of S with the number of stages. For example, if the number of stages is increased from 1 to 5, the figure of merit increases only by the factor 3.66, while for a further increase in n the gain is very slow.[5] This gain, furthermore, is accompanied by a serious increase in the "sluggishness" of the circuit: That is, in the partciular example cited, by an increase of 5 to 1 in the time required for signals to build up to their steady state.

The analysis of a number of representative schemes, such as the introduction of resistance to damp out disturbances, balancing schemes designed to neutralize static without affecting the signal, detuning to change the natural oscillation frequency of the circuit, demodulation through several frequency stages, etc., has shown that they are one and all without value in increasing the ratio of mean square signal to interference current. In the light of the general theory, the reason for this is clear and the limitation imposed on the solution of the static problem by means of selective circuits is seen to be inherent in the nature of the interference itself.

[4] See U. S. Patent No. 1173079 to Alexanderson.

[5] When the number of stages n is fairly large, the selective figure of merit becomes proportional to $\sqrt{n}$ and the building-up time to n.

Some Contemporary Advances in Physics—VII
Waves and Quanta

By KARL K. DARROW

THE invaluable agent of our best knowledge of the environing world, and yet itself unknown except by inference; the intermediary between matter and the finest of our senses, and yet itself not material; intangible, and yet able to press, to strike blows, and to recoil; impalpable, and yet the vehicle of the energies that flow to the earth from the sun—light in all times has been a recognized and conspicuous feature of the physical world, a perpetual reminder that the material, the tangible, the palpable substances are not the only real ones. Yet its apparent importance, to our forerunners who knew only the rays to which the eye responds and suspected no others, was as nothing beside its real importance, which was realized very gradually during the nineteenth century, as new families of rays were discovered one after the other with new detecting instruments and with new sources. Radiation is not absent from the places where there is no eye-stimulating light; radiation is omnipresent; there is no region of space enclosed or boundless, vacuous or occupied by matter, which is not pervaded by rays; there is no substance which is not perpetually absorbing rays and giving others out, in a continual interchange of energy, which either is an equilibrium of equal and opposite exchanges, or is striving towards such an equilibrium. Radiation is one of the great general entities of the physical world; if we could still use the word "element," not to mean one of the eighty or ninety kinds of material atoms, but in a deeper sense and somewhat as the ancients used it, we might describe radiation and matter, or possibly radiation and electricity, as coequal elements. Also the problem of the nature and structure of radiation is of no lesser importance than the problem of the structure and nature of matter; and in fact neither can be treated separately; they are so inextricably intertwined that whoever sets out to expound the present condition of one soon finds himself outlining the other. One cannot write a discourse on the nature of radiation alone nor on the structure of the atom alone, one can but vary the relative emphasis laid upon these two subjects, or rather upon these two aspects of a single subject; and in this article I shall restate many things about the atom which were stated in former articles, but the emphasis will be laid upon light.

Speaking very generally and rather vaguely, light has been much more tractable to the theorists than most of the other objects of enquiry in physics or chemistry. Over a rather long period of years, it was indeed generally regarded as perfectly intelligible. The famous battle between the corpuscular theory adopted by Newton, and the wave-theory founded by Descartes and Huyghens, died out in the earlier years of the nineteenth century with the gradual extinction of the former. The history of optics in the nineteenth century, from Fresnel and Young to Michelson and Rayleigh, is the tale of a brilliant series of beautiful and striking demonstrations of the wave-theory, of experiments which were founded upon the wave-theory as their basis and would have failed if the basis had not been firm, of instruments which were designed and competent to make difficult and delicate measurements of all sorts—from the thickness of a sheet of molecules to the diameter of a star—and would have been useless had the theory been fallacious. The details of the bending of light around the sides of a slit or the edge of a screen, the intricate pattern of light and shade formed where subdivisions of a beam of light are reunited after separation, the complexities of refraction through a curved surface, are represented by the theory with all verifiable accuracy; and so are the incredibly complicated phenomena attending the progress of light through crystals, phenomena which have slipped out of common knowledge because few are willing to undertake the labour of mastering the theory. The wave-theory of light stands with Newton's inverse-square law of gravitation, in respect of the many extraordinarily precise tests which it has undergone with triumph; I know of no other which can rival either of them in this regard.

By the term "wave-theory of light" I have meant, in the foregoing paragraph, the conception that light is a wave-motion, an undulation, a periodic form advancing through space without distorting its shape; I have not meant to imply any particular answer to the question, *what is it of which light is a wave-motion?* It may seem surprising that one can make and defend the conception, without having answered the question beforehand; but as a matter of fact there are certain properties common to all undulations, and these are the properties which have been verified in the experiments on light. There are also certain properties which are not shared by such waves as those of sound, in which the vibration is confined to a single direction (that normal to the wavefront) and may not vary otherwise than in amplitude and phase, but are shared by transverse or distortional waves in elastic solids, in which the vibration may lie in any of an infinity of directions (any direction tangent to the wavefront). Light possesses these properties, and therefore the wave-motion which is radiation may not be compared with the wave-motion which is sound; but a wide range of comparisons still remains open.

Of course, very many have proposed images and models for "the thing of which the vibrations are light", and many have believed with an unshakable faith in the reality of their models. The fact that light-waves may be compared, detail by detail, with transverse vibrations in an elastic solid, led some to fill universal space with a solid elastic medium to which they gave the sonorous name of "luminiferous aether". It is not many years since men of science used to amaze the laity with the remarkable conception of a solid substance, millions of times more rigid than steel and billions of times rarer than air, through which men and planets serenely pass as if it were not there. Even now one finds this doctrine occasionally set forth.[1]

In that image of the elastic solid, the propagation of light was conceived to occur because, when one particle of the solid is drawn aside from its normal place, it pulls the next one aside, that one the next one to it, and so on indefinitely. Meanwhile, each particle which is drawn aside exerts a restoring force upon the particle of which the displacement preceded and caused its own. Set one of the particles into vibration, and the others enter consecutively into vibration. Maintain the first particle in regular oscillation, and each of the others oscillates regularly, with a phase which changes from one to the next; a wave-train travels across the medium. One particle influences the next, because of the attraction between them. But in the great and magnificent theory of light which Maxwell erected upon the base of Faraday's experiments, the propagation was explained in an altogether different manner. Vary the magnetic field across a loop of wire in a periodic manner, and you obtain a periodic electric force around the loop, as is known to everyone who has dabbled in electricity. Vary the electric field periodically, and you obtain a periodic magnetic field—this a fact not by any means so well known as the other, one which it was Maxwell's distinction to have anticipated, and which was verified after the event. In a traveling train of light-waves the electric field and the magnetic field stimulate one another alternately and reciprocally, and for this reason the wave-train travels. Since the periodic electric field may point in any one of the infinity of directions in the plane of the wave-front, the wave-motion possesses all the freedom and variability of

[1] Apparently the image of the elastic solid was never quite perfected; one recalls the question as to whether its vibrations were in or normal to the plane of polarization of the light, which required one answer in order to agree with the phenomena of reflection, and another in order to agree with those of double refraction. Probably a *modus vivendi* could have been arranged if the whole idea had not been superseded.

form which are required to account for the observed properties of light.

Maxwell's theory immediately achieved the stunning success of presenting a value for the speed of the imagined electromagnetic waves, determined exclusively from measurements upon the magnetic fields of electric currents, and agreeing precisely with the observed speed of light. Two supposedly distinct provinces of physics, each of which had been organized on its own particular basis of experience and in its own particular manner, were suddenly united by a stroke of synthesis to which few if any parallels can be found in the history of thought. And this is by no means the only achievement of the electromagnetic theory of light; there will shortly be occasion to mention some of the others.

Now that there was so much evidence that light travels as a wave-motion, and that its speed and other properties are those of electromagnetic waves, it became urgently desirable to inquire into the nature of the *sources* of light. Granted that light *en route* outwards from a luminous particle of matter is of the nature of a combination of wave-trains, what is taking place in the luminous particle? To this question all our experience and all our habits of thought suggest one sole obvious answer—that in the luminous particle there is a vibrating something, a *vibrator*, or more likely an enormous number of vibrators—one to each atom, possibly—and the oscillations of these vibrators are the sources of the waves of light, as the oscillations of a violin-string or a tuning-fork are the sources of waves of sound. This analogy drawn from acoustics, this picture of the vibrating violin-string and the vibrating tuning-fork, has been powerful—indeed, it begins to seem, too powerful—in guiding the formation of our ideas on light. It is profitable to reflect that the evolution of thought in acoustics must have traveled in the opposite sense from the evolution of thought in optics. Whoever it was who was the first to conceive that sound is a wave-motion in air, must certainly have arrived at the idea by noticing that sounding bodies vibrate. One feels the trembling of the tuning-fork or the bell, one sees the violin-string apparently spread out into a band by the amplitude of its motion; it is not difficult to build apparatus which, like a slowed-down cinema film, makes the vibrations separately visible, or, like the stroboscope, produces an equivalent and not misleading illusion. This was not possible in optics, and never will be. In acoustics, one may sometimes accept the vibrations of the sounding body as an independently-given fact of experience, and reason forward to the wave-motion spreading outwards into the environing air; in optics, this entrance to the path is closed, one must reason in the inverse sense from the wave-motion to the qualities of the shining body. Inevitably, it was assumed that when the path should at last be successfully retraced, the shining body would be found in the semblance of a vibrator.

For a few years at the end of the nineteenth century and the beginning of the twentieth, it seemed that the desired vibrator had been found. Apparently it was the electron, the little corpuscle of negative electricity, of which the charge and the mass were rather roughly estimated in the late nineties, although Millikan's definite measurements were not to come for a decade yet. Maxwell had not conceived of particles of electricity, his conception of the "electric fluid" was indeed so sublimated and highly formal that it gave point to the celebrated jest (I think a French one) about the man who read the whole of his "Electricity and Magnetism" and understood it all except that he was never able to find out what an electrified body was. H. A. Lorentz incorporated the electron into Maxwell's theory. Conceiving it as a spherule of negative electricity, and assuming that in an atom one or more of these spherules are held in equilibrium-positions, to which restoring-forces varying proportionally to displacement draw them back when they are displaced, Lorentz showed that these "bound" electrons are remarkably well adapted to serve as sources and as absorbents for electromagnetic radiation. Displaced from its position of equilibrium by some transitory impulse, and then left to itself, the bound electron would execute damped oscillations in one dimension or in two, emitting radiation of the desired kind at a calculable rate. Or, if a beam of radiation streamed over an atom containing a bound electron, there would be a "resonance" like an acoustic resonance—the bound electron would vibrate in tune with the radiation, absorbing energy from the beam and scattering it in all directions, or quite conceivably delivering it over in some way or other to its atom or the environing atoms. There were numerical agreements between this theory and experience, some of them very striking.[2] Apparently the one thing still needful was to produce a plausible theory of these binding-forces which control the response of the "bound" electron to disturbances of all kinds. Once these were properly described, the waves of light would be supplied with

[2] Notably, the trend of the dispersion-curves for certain transparent substances, recently extended by Bergen Davis and his collaborators to the range of X-ray frequencies; the normal Zeeman effect; Wien's observations on the exponential dying-down of the luminosity of a canal-ray beam, interpreted as the exponential decline in the vibration-amplitudes of the bound electrons in the flying atoms; the dependence of X-ray scattering on the number of electrons in the atom.

their vibrators, the electromagnetic theory would receive a most valuable supplement. And, much as a competent theory of the binding-forces was to be desired, a continuing failure to produce one would not impugn the electromagnetic theory, which in itself was a coherent system, self-sustaining and self-sufficient.

This was the state of affairs in the late nineties. The wave-conception of light had existed for more than two centuries, and it was seventy-five years since any noticeable opposition had been raised against it. The electromagnetic theory of light had existed for about thirty years, and now that the electron had been discovered to serve as a source for the waves which in their propagation through space had already been so abundantly explained, there was no effective opposition to it. Not all the facts of emission and absorption had been accounted for, but there was no reason to believe that any particular one of them was unaccountable. Authoritative people thought that the epoch of great discoveries in physics was ended. It was only beginning.

In the year 1900, Max Planck published the result of a long series of researches on the character of the radiation inside a completely-enclosed or nearly-enclosed cavity, surrounded by walls maintained at an even temperature. Every point within such a cavity is traversed by rays of a wide range of wave lengths, moving in all directions. By the "character" of the radiation, I mean the absolute intensities of the rays of all the various frequencies, traversing such a point. The character of the radiation, in this sense, is perfectly determinate; experiment shows that it depends only on the temperature of the walls of the cavity, not on its material. According to the electromagnetic theory of radiation, as completed by the adoption of the electron, the walls of the cavity are densely crowded with bound electrons; nor are these electrons all bound in the same manner, so that they would all have the same natural frequency of oscillation—they are bound in all sorts of different ways with all magnitudes of restoring-forces, so that every natural frequency of oscillation over a wide range is abundantly represented among them. Now the conclusion of Planck's long study was this:

If the bound electrons in the walls of the cavity (i.e., *in any solid body*) *did really radiate while and as they oscillate, in the fashion prescribed by the electromagnetic theory, then the character of the radiation in the cavity would be totally different from that which is observed.*[3]

[3] The belief that the character of radiation within a cavity could not be explained without doing some violence to the "classical mechanics" had already been gaining ground for some years, by reason of extremely recondite speculations of a statistical nature. It is very difficult to gauge the exact force and bearing of such considerations.

However, *if the bound electrons do not radiate energy while they oscillate, but accumulate it and save it up and finally discharge it in a single outburst when it attains some one of a certain series of values* $h\nu$, $2h\nu$, $3h\nu$, etc. (h stands for a constant factor, ν for the frequency of vibration of the electrons and the emitted radiation)—*then the character of the radiation will agree with that which is observed*, provided a suitable value be chosen for the constant h.

The value required [4] for h in C.G.S. units (erg seconds) is $6.53 . 10^{-27}$.

Here, then, was a phenomenon which the electromagnetic theory seemed to be fundamentally incapable of explaining. For this notion of a bound electron, which oscillates and does not meanwhile radiate, is not merely foreign to the classical theory, but very dangerous to it; one does not see how to introduce it, and displace the opposed notion, without bringing down large portions of the structure (including the numerical agreements which I cited in a foregoing footnote). However, Planck had arrived at this conclusion by an intricate process of statistical and thermodynamical reasoning. Statistical reasoning is notoriously the most laborious and perplexing in all physics, and many will agree that thermodynamical reasoning is not much less so. Planck's inference made an immense impression on the most capable thinkers of the time; but in spite of the early adherence of such men as Einstein and Poincaré, I suspect that even to this day it might practically be confined to the pages of the more profound treatises on the philosophical aspects of physics, if certain experimenters had not been guided to seek and to discover phenomena so simple that none could fail to apprehend them, so extraordinary that none could fail to be amazed.

Honour for this guidance belongs chiefly to Einstein. Where Planck in 1900 had said simply that bound electrons emit and absorb energy in fixed finite quantities, and shortly afterwards had softened his novel idea as far as possible by making it apply only to the act of emission, Einstein in 1905 rushed boldly in and presented the idea that these fixed finite quantities of radiant energy retain their identity throughout their wanderings through space from the moment of emission to the moment of absorption. This idea he offered as a "heuristic" one—the word, if I grasp its connotation exactly, is an apologetic sort of a word, used to describe a theory which achieves successes though its author feels at heart that it really is too absurd to

[4] I take the numerical values of the constant h scattered through this article from Gerlach. The weighted mean of the experimental values, with due regard to the relative reliability of the various methods, is taken as 6.55 or $6.56 . 10^{-27}$. None of the individual values cited in these pages is definitely known to differ from this average by more than the experimental error.

be presentable. The implication is, that the experimenters should proceed to verify the predictions based upon the idea, quite as if it were acceptable, while remembering always that it is absurd. If the successes continue to mount up, the absurdity may be confidently expected to fade gradually out of the public mind. Such was the destiny of this heuristic idea.

I will now describe some of these wonderfully simple phenomena — wonderfully simple indeed, for they stand out in full simplicity in domains where the classical electromagnetic theory would almost or quite certainly impose a serious complexity. If Planck's inference from the character of the radiation within a cavity had been deferred for another fifteen years, one or more of these phenomena would assuredly have been discovered independently. What would have happened in that case, what course the evolution of theoretical physics would have followed, it is interesting to conjecture.

The *photoelectric effect* is the outflowing of electrons from a metal, occurring when and because the metal is illuminated. It was discovered by Hertz in 1889, but several years elapsed before it was known to be an efflux of electrons, and several more before the electrons were proved to come forth with speeds which vary from one electron to another, upwards as far as a certain definite maximum value, and never beyond it.

Here is a rather delicate point of interpretation, which it is well to examine with some care; for all the controversies as to continuity versus discontinuity in Nature turn upon it, in the last analysis. What is meant, or what reasonable thing can be meant, when one says that the speeds of all the electrons of a certain group are confined within a certain range, extending up to a certain limiting topmost value? If one could detect each and every electron separately, and separately measure its speed, the meaning would be perfectly clear. For that matter, the statement would degenerate into a truism. The fact is otherwise. The instruments used in work such as this perceive electrons only in great multitudes. Suppose that one intercepts a stream of electrons with a metal plate connected by a wire to an electrometer. If a barrier is placed before the electrons in the form of a retarding potential-drop, which is raised higher and higher, the moment eventually comes when the current into the electrometer declines. This happens because the slower electrons are stopped and driven back before they reach the plate, the faster ones surmount the barrier. As the potential-drop is further magnified, the reading of the electrometer decreases steadily, and at last becomes inappreciable. Beyond a certain critical value of the retarding voltage, the electrometer reports no influx of electrons. Does this really mean that there are *no* electrons with more than just the speed necessary to overpass a retarding voltage of just that critical value? Or does it merely mean that the electrons flying with more than that critical speed are plentiful, but not quite plentiful enough to make an impression on the electrometer? Is there any topmost speed at

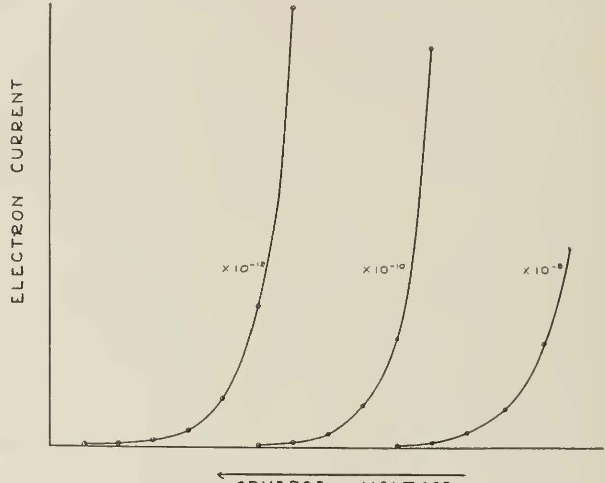


Fig. 1—Curves showing thermionic electron-current versus opposing voltage, demonstrating a distribution-in-speed extending over an unlimited range of speeds. Multiply the ordinates of the middle curve by 100, those of the right-hand curve by 10,000, to bring them to the same scale and make them merge into a single curve. (L. H. Germer)

all, or should we find, if we could replace the current-measuring device with other and progressively better ones *ad infinitum*, that the apparent maximum speed soared indefinitely upwards?

Absolute decisions cannot be rendered in a question of this kind; but it is possible, under the best of circumstances, to pile up indicatory evidence to such an extent that only an unusually strong will-to-disbelieve would refuse to be swayed by it. The judgment depends on the shape of the curve which is obtained by plotting the electrometer-reading *vs.* the retarding potential—in other words, the fraction y of the electrons of which the energy of motion surpasses the amount x, determined from the retarding-voltage by the relation $x=eV$. Look for example at the curves of Fig. 1, which refer to the electron-

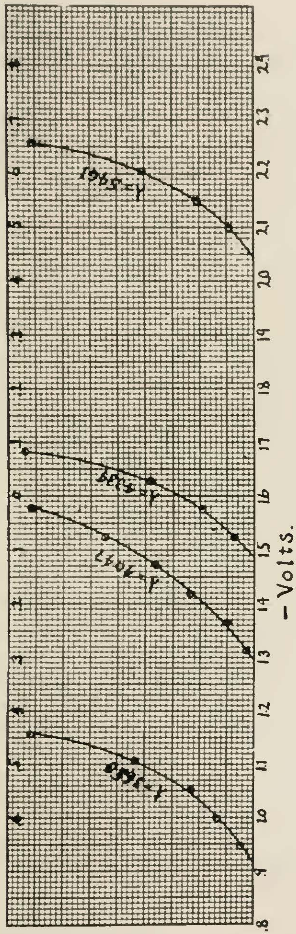


Fig. 2 Curves showing photoelectric electron-current versus opposing voltage, demonstrating a distribution-in-speed extending over a range limited at the top. (R. A. Millikan, *Physical Review*)

stream flowing spontaneously out of an incandescent wire; they are three segments of one single curve, plotted on different scales as the numerals show. This curve bends so gradually around towards tangency with the axis of abscissae, that one can hardly avoid the inference that it is really approaching that axis as if to an asymptote, and that if the electrometer at any point ceases to declare a current, it is because the electrometer is too insensitive to respond to the smaller currents, and not because there are no faster electrons. Look instead at the curves of Fig. 2, which refer to the electrons emerging from an illuminated surface of sodium. These curves slant so sharply towards the axis of abscissae, they bend so slightly in the portions of their courses where the data of experiment determine them, that the linear extrapolation over the little interval into the axis commends itself as natural and inevitable. Because the curves for the thermionic electrons approach the axis so gently, it is agreed that their velocities are distributed continuously over an unlimited range; because the curves for the photoelectrons cut into it so acutely, it is felt that their velocities are confined below a definite maximum value.

This therefore is the photoelectric effect: waves of light inundate the surface of a metal, and electrons pour out with various velocities, some nearly attaining and none exceeding a particular topmost value. I will designate this maximum speed, or rather the corresponding maximum kinetic energy, by E_{max}. Analyzing the process in the classical manner, one must imagine the waves entering into the metal and setting the indwelling electrons into forced oscillations; the oscillations grow steadily wider; the speed with which the electron dashes through its middle position grows larger and larger, and at last it is torn from its moorings and forces its way through the surface of the metal. Some of the energy it absorbed during the oscillations is spent (converted into potential energy) during the escape; the rest is the kinetic energy with which it flies away. Even if the electron were free within the metal and could oscillate in response to the waves, unrestrained by any restoring force, it would still have to spend some of its acquired energy in passing out through the boundary of the metal (the laws of thermionic emission furnish evidence enough for this). It is natural to infer that E_{max} is the energy absorbed by an electron originally free, minus this amount (let me call it P) which it must sacrifice in crossing the frontier; the electrons which emerge with energies lower than E_{max} may be supposed to have made the same sacrifice at the frontier and others in addition, whether in tearing themselves away from an additional restraint or in colliding with atoms during their emigration. This is not the only conceivable interpretation, but it seems unprofitable to enter into the others. It is therefore E_{max} which appears to merit the most attention.

Now the mere fact that there is a maximum velocity of the escaped electrons, that there is an E_{max}, is not in itself of a nature to suggest that the classical theory is inadequate. It is the peculiar dependence of this quantity on the two most important controllable qualities of the light—on its intensity and on its frequency—which awakens the first faint suspicions that something has at last been discovered, which the classical theory is ill adapted to explain. One would predict with a good deal of confidence that the greater the intensity of the light, the greater the energy acquired by the electron in each cycle of its forced oscillation would be, the greater the energy with which it would finally break away, the greater the residuum of energy which at the end would be left to it. But E_{max} is found to be independent of the intensity of the light. This is strange; it is as though the waves beating upon a beach were doubled in their height and the powerful new waves disturbed four times as many pebbles as before, but did not displace a single one of them any farther nor agitate it any more violently than the original gentle waves did to the pebbles that they washed about. As for the dependence of E_{max} on the frequency of the light, it would be necessary to make additional assumptions to calculate it from the classical theory; in any case it would probably not be very simple. But the actual relation between E_{max} and ν is the simplest of all relations, short of an absolute proportionality; this is it:

$$E_{max} = h\nu - P \tag{1}$$

Fig. 3 shows the relation for sodium, observed by Millikan.

The maximum energy of the photoelectrons increases linearly with the frequency of the light. P is a constant which varies from one metal to another. In the terms of the simple foregoing interpretation, P is the energy which an electron must spend (more precisely, the energy which it must invest or convert into potential energy) when it passes through the frontier of the metal on its way outward. Comparing the values of P for several metals with the contact potentials which they display relatively to one another, one finds powerful evidence confirming this theory. Having discussed this particular aspect of the question in the fifth article of this series, I will not enter further into it at this point.

The constant h is the same for all the metals which have been used in such experiments. The best determinations have been made upon two or three of the alkali metals, for these are the only metals

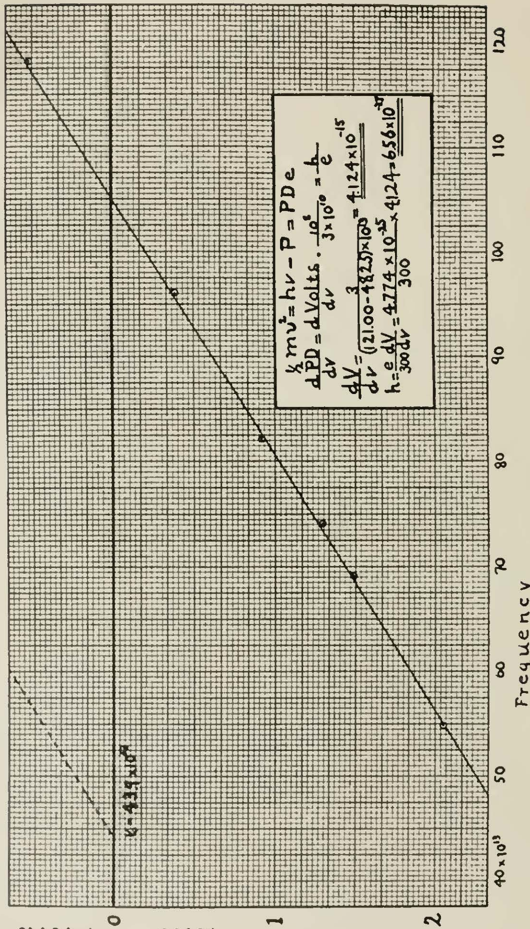


Fig. 3—Curve showing the linear relation between the maximum energy of photo-electrons and the frequency of the light which excites them. (Millikan, *Physical Review*)

which release electrons when illuminated with light of wide convenient ranges of frequency and color. Most metals must be irradiated with ultraviolet light, and the experiments become very difficult if they must be performed with light of frequencies far from the visible spectrum. The values which Millikan obtained for sodium and for lithium agree within the experimental error with one another and with the mean value

$$h = 6.57 \cdot 10^{-27} \tag{2}$$

The maximum energy of the electrons released by light of the frequency ν is therefore equal to a quantity $h\nu$ which is the same, whatever metal be illuminated by the light—a quantity which is characteristic of the light, not of the metal—minus a quantity P which, there is every reason to believe, is the quota of energy surrendered by each electron in passing out across the boundary-surface of the metal. It is *as if* each of the released electrons had received a quantity $h\nu$ of energy from the light. I will go one step further, and lay down this as a rule, with another cautiously-inserted *as if* to guard against too suddenly daring an innovation:

Photoelectric emission occurs as if *the energy in the light were concentrated in packets, or units, or corpuscles of amount $h\nu$, and one whole unit were delivered over to each electron.*

This is a perfectly legitimate phrasing of equation (1), but I doubt whether anyone would ever have employed it, even with the guarded and apologetic *as if*, but for the fact that the value of h given in (2) agreed admirably well with the value of that constant factor involved in Planck's theory, the constant to which he had given this very symbol and a somewhat similar role. Deferring for a few pages one other extremely relevant feature of the photoelectric effect (its "instantaneity") I will proceed to examine these other situations.

An effect which might well be, though it is not, called the *inverse photoelectric effect*, occurs when electrons strike violently against metal surfaces. Since radiation striking a metal may elicit electrons, it is not surprising that electrons bombarding a metal should excite radiation. Electrons moving as slowly as those which ultraviolet or blue light excites from sodium do not have this power; or possibly they do, but the radiation they excite is generally too feeble to be detected. Electrons moving with speeds corresponding to kinetic energies of hundreds of equivalent volts,[5] and especially electrons

[5] One equivalent volt of energy = the energy acquired by an electron in passing across a potential-rise of one volt $= e/300$ ergs $= 1.591 \cdot 10^{-12}$ ergs. This unit is usually called simply a "volt of energy", or "volt", a bad usage but ineradicable. Also "speed" is used interchangeably with "energy" in speaking of electrons, and one finds (and, what is worse, cannot avoid) such deplorable phrases as "a speed of 4.9 volts" ! ! !

with energies amounting to tens of thousands of equivalent volts, do possess it. This is in fact the process of excitation of X-rays, which are radiated from a metal target exposed to an intense bombardment of fast electrons. The protagonists of the electromagnetic theory had an explanation ready for this effect, as soon as it was discovered. A fast electron, colliding with a metal plate, is brought to rest by a slowing-down process, which might be gradual or abrupt, uniform or *saccadé*, but in any case must be continuous. Slowing-down entails radiation; the radiation is not oscillatory, for the electron is not oscillating, but it is radiation none the less; it is an outward-spreading single pulsation or *pulse*, comparable to the narrow spherical shell of condensed air which diverges outward through the atmosphere from an electric spark and has been photographed so often, or to a transient in an electrical circuit.

One may object that the pulse is just a pulse and nothing more, while the X-rays are wave-trains, for otherwise the X-ray spectroscope (which is a diffraction apparatus) would not function. The objection is answered by pointing out the quite indubitable fact that any pulse, whatever its shape (by "shape" I mean the shape of the curve representing the electric field strength, or whatever other variable one chooses to take, as a function of time at a point traversed by the wave) can be accurately reproduced by superposing an infinity of wave-trains, of all frequencies and divers properly-adjusted amplitudes, which efface one another's periodic variations, and in fact efface one another altogether at all moments except during the time-interval while the pulse is passing over—during this interval they coalesce into the pulse. Thence, the argument leads to the contention that the actual pulse is made up of just such wave-trains, and the sapient diffracting crystal recognizes them all and diffracts each of them duly along its proper path. The problem is not new, nor the answer; white light has long been diagnosed as consisting of just such pulses, and the method of analyzing transient impulses in electrical circuits into their equivalent sums of wave-trains has been strikingly successful.

The application of the method to this case of X-ray excitation enjoyed one qualitative success. The spherical pulse diverging from the place where an electron was brought to rest should not be of equal thickness at all the points of its surface; it should be broader and flatter on the side towards the direction whence the electron came, thinner and sharper on the side towards the direction in which the electron was going when it was arrested. Analyzing the pulse, it is found that at the point where it is broad and low, the most intense of its equivalent wave-trains are on the whole of a lower frequency than the most intense of the wave-trains which constitute it where it is narrow and high. By examining and resolving the X-rays radiated from a target, at various inclinations to the direction of the bombarding electrons, this was verified—verified in part, not altogether. The X-rays radiated nearly towards the source of the electron-stream include a

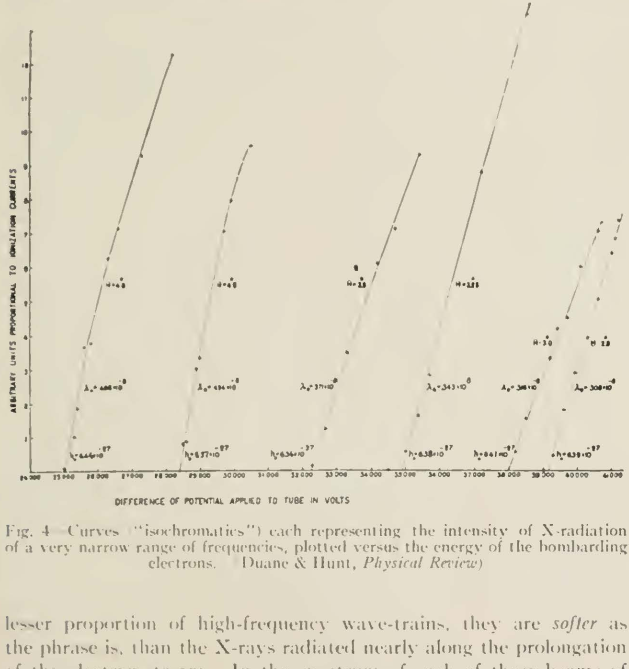


Fig. 4. (Curves ("isochromatics") each representing the intensity of X-radiation of a very narrow range of frequencies, plotted versus the energy of the bombarding electrons. (Duane & Hunt, *Physical Review*)

lesser proportion of high-frequency wave-trains, they are *softer* as the phrase is, than the X-rays radiated nearly along the prolongation of the electron-stream. In the spectrum of each of these beams of X-rays, there is a wave length where the density of radiant energy attains a maximum, and this wave length is longer in the former beam than in the latter one. So much is implied in the classical theory.

But it is nowhere implied in the classical theory that the spectrum of an X-ray beam, produced when electrons of a constant energy rain down upon a metal, should extend upwards only to a certain maximum frequency, and then and there come to a sudden end; yet apparently it does. There is a ***high-frequency limit*** to each X-ray spectrum, and wave-trains of frequencies exceeding that limit are not detected; whereas the spectrum of the hypothetical pulses ought to include wave-trains of every frequency low or high, the amplitudes indeed declining to infinitely low values as one goes along the spectrum to infinitely high frequencies, but certainly declining smoothly and gradually. To demonstrate this high-frequency limit is a delicate experimental problem, quite like that other problem of demonstrating a sharply definite topmost value for the energies of photoelectrons. That question whether the curves of photoelectric current *vs.* retarding voltage, the curves of Fig. 2, cut straightly and sharply enough into the axis of abscissae to prove that there are no photoelectrons with velocities higher than the one corresponding to x_o, returns again in a slightly altered form.

The most reliable of the methods actually used to demonstrate the high-frequency limit depends on the fact that the high limiting frequency (which I will call $\nu_{\max}$) varies with the energy of the bombarding electrons, increasing as their velocity increases. Therefore, if the radiant energy belonging to rays of a certain fixed wave length or a certain fixed narrow range of wave lengths is separated out from the X-ray beam by a spectroscope, and measured for various velocities of the impinging electrons, passing from very high velocities step by step to very low ones; it will decrease from its first high value to zero at some intermediate velocity, and thereafter remain zero. But according to the classical theory also, it must decrease from its first high value to an imperceptibly low one; the descent however will be gradual and smooth. Thus the only question which can be settled by experiment is the question whether the descent from measurable intensities to immeasurably small ones resembles the gentle quasi-asymptotic decline of the curve of Fig. 1 or the precipitate slope of the curve of Fig. 2. The data assembled by Duane and Hunt are shown in Fig. 4 plotted in the manner I have described; there is little occasion for doubt as to which sort of curve these resemble most.[6]

Each of the curves in Fig. 4 represents that portion of the total intensity of an X-ray beam, which belongs to rays of wave lengths near the marked value of the frequency ν. This frequency is the high

[6] Three simple curves of the intensity-distribution in the X-ray spectrum are shown in Figure 5. The abscissa is neither frequency or wavelength, but a variable which varies continuously with either (it is actually *arc sin* of a quantity proportional to wavelength) so that the acute angle between each curve and the axis of abscissae, at the point where they meet, corresponds to and has much the same meaning as the acute angles in Figure 2—not so conspicuously.

limiting frequency ν_{max} for that value of the energy E of the bombarding electrons, which corresponds to the point on the axis of abscissae where the curve (extrapolated) intersects it. The relation between ν_{max} and E is the simplest of all relations:

$$E = \text{constant} \cdot \nu_{max} = h\ \nu_{max} \tag{3}$$

The constant h is the same for all the metals on which the experiment has been performed—a few of the least fusible ones, for metals of a low melting-point would be melted before E could be lifted far enough to give an adequate range for determining the relation between it and

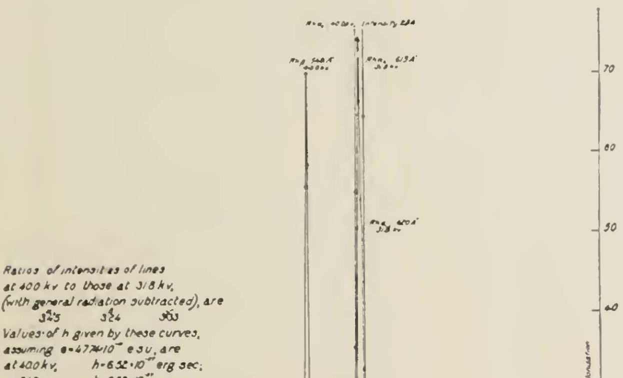


Fig. 5—The continuous X-ray spectrum for three values of the energy of the bombarding electrons, intensity being plotted versus a quantity varying uniformly with frequency. Ignore the peaks. (D. L. Webster, *Physical Review*.) See footnote 6

ν_{max}. The value[7] given for it by Gerlach, after a critical study of all the determinations, is

$$h = 6.53.10^{-27} \tag{4}$$

The highest frequency of radiation which electrons moving with the energy E are able to excite, when they are brought to rest by colliding with a metal target, is therefore equal to E divided by a constant independent of the kind of metal. So far as this high limiting frequency is concerned, it is perfectly legitimate to express equation (3) in these words,

Excitation of radiation by electrons stopped in their flight by collision with a metal occurs as if *the energy in the radiation were concentrated in units of amount $h\nu$, and one such unit were created out of the total energy which each electron surrenders when it is stopped.*

As for the radiation of frequencies inferior to the high limiting frequency, it is very easily explained by asserting that most of the electrons come to rest not in one operation, but in several successive ones, dividing their energy up among several units of frequencies inferior to ν_{max} or E/h; or possibly they lose energy in various sorts of impacts or various other ways before making the first impact of the sort which transforms their energy into energy of X-rays. Nothing about it contradicts the italicized rule. Still it is not likely that anyone would have formulated equation (3) in such language, if the value of the constant h which appears in it were not identical with the value which we have already once encountered in analyzing the photoelectric effect, and with the value at which Planck earlier arrived.

I think it is too early in this discourse to fuse these italicized Rules for the release of electrons by radiation and the excitation of radiation by electrons into a single Rule; but by contemplating the two Rules side by side one arrives without much labor at an inference which could be tested even though we had no way of measuring the frequency of a radiation, and in fact was verified before any such way existed. For if electrons of energy E can excite radiation of frequency E/h, and radiation of frequency E/h striking a piece of metal can elicit electrons of energy $h(E/h) - P$; then, if a target is bombarded with electrons, and another metal target is exposed to the radiation which emanates from the first one, the fastest of the electrons which escape from the second target will move with the same velocity and

[7] Gerlach regards this as the most accurate of all the methods for determining h, an opinion in which probably not all would concur. It has been maintained that the high-frequency limit, like the wavelength of maximum intensity in the X-ray spectrum, depends on the inclination of the X-ray beam to the exciting electron-stream. I do not know whether the experiments adduced in support of this claim have been adequately confuted.

the same energy as the electrons which strike the first one (minus the quantity P which, however, is immeasurably small and perfectly negligible in comparison with the energy of the electrons which excite ordinary X-rays). This fact emerged from a series of experiments which were performed by various people in the first decade of this century, the results of which were generally phrased somewhat in this way, "the energy of the secondary electrons depends only on the energy of the primary electrons, not on the nature of the material which the primary electrons strike or on that from which the secondary electrons issue, nor on the distance over which the X-rays travel." Upon these results Sir William Bragg based his corpuscular theory of X-rays; for (he argued) the most sensible interpretation of the facts is surely this, that some of the electrons striking the first target rebound with their full energy, and rebound again with their full energy from the second target, each of them carrying with it from the first to the second target a positive particle which neutralizes its charge over that part of its course, and so defeats all the methods devised to recognize a flying electron. Not many years later, Sir William cooperated in the slaying of his own theory, by developing the best of all methods for proving that X-rays are undulatory and measuring their wave-lengths; but it was only the imagery of the theory that perished, for its essence, the idea that the energy of the first electron travels as a unit or is carried as a parcel to the place where the second electron picks it up, had to be resurrected. All the mystery of the contrast between wave-theory and quantum-theory is implicit in this phenomenon, for which Sir William found an inimitable simile: "It is as if one dropped a plank into the sea from a height of 100 feet, and found that the spreading ripple was able, after travelling 1,000 miles and becoming infinitesimal in comparison with its original amount, to act upon a wooden ship in such a way that a plank of that ship flew out of its place to a height of 100 feet."

Among the radiations excited from a metal by electrons of a single energy E, there are many of which the frequencies differ from the interpreted frequency E/h, being lower. Among the electrons expelled from a metal by radiation of a single frequency ν, there are many of which the energies differ from the interpreted energy-value $h\nu$, being lower. These were accounted for by supposing that the electrons are troubled by repeated encounters with closely-crowded atoms. If then a metal vapor or a gas were bombarded with electrons or exposed to radiation, would all the excited radiation have a single frequency conforming to equation (3), would all the released electrons have a single energy conforming to equation (1)? One could not affirm this *a priori*, for a solid metal is not a collection of free atoms close together as a gas is an assemblage of free atoms far apart, but rather a structure of atoms which interfere with one another and are distorted, and there are many electrons in a solid of which the bonds and the constraints are very different from those by which the electrons of free atoms are controlled and vice versa. When a plate of sodium or a pool of mercury is exposed to a rain of electrons, not exceeding say 10 equivalent volts in energy, nothing apparent happens.[8] When the vapor of either metal is similarly exposed, the atoms respond in a manner from which they are inhibited, when they are bound together in the tight latticework of a solid or the promiscuous crowding of a liquid; and light is emitted.

The phenomena are clearest when the bombarded vapor is that of a volatile metal, such as mercury, sodium, or magnesium. The atoms in such vapors are not usually bound together two by two or in greater clusters, as they are in such gases as oxygen or hydrogen, of which the response to electron-impacts or to radiation is not quite understood to this day; and the first radiations which they emit are not in the almost inaccessible far ultra-violet, like those of the monatomic noble gases, but in the near ultra-violet or even in the visible spectrum. Dealing with such a vapor, I will say mercury for definiteness, one observes that so long as the energy of the bombarding electrons remains below a certain value, no perceptible light is emitted; but beyond, there is a certain range of energies, such that electrons possessing them are able to arouse one single frequency of radiation from the atoms. Ordinarily, as when a vapor is kept continuously excited by a self-sustaining electric discharge throughout it, the atoms emit a great multitude of different frequencies of radiation, forming a rich and complicated spectrum of many lines. But if the energy of the bombarding electrons is carefully adjusted to some value within the specified range, only one line of this spectrum makes its appearance; under the best of circumstances this single line may be exceedingly bright, so that the absence of its companions—some of which, in an ordinary arc-spectrum, are not much inferior to it in brightness—is decidedly striking. The one line which constitutes this *single-line spectrum* is the first line of the principal series in the complete arc-spectrum of the element; its wave length is (to take a few examples) 2536A for mercury, 5890 for sodium (for which it is a doublet), 4571 for magnesium.

[8] According to a very recent paper by C. H. Thomas, radiations from iron excited by electrons with as low an energy as some two or three equivalent volts have been detected.

Does this single line appear suddenly at a precise value of the energy of the impinging electrons? This question suggests itself, when one has already studied the excitation of X-rays from solids by electrons and the excitation of electrons from solids by light. Here again we meet that tiresome but ineluctable problem, as to what constitutes a *sudden* appearance, and how we should recognize it if it really occurred. The only consistent way to meet it (consistent, that is, with the ways already employed in the prior cases) would be to measure the intensity of the line for various values of the energy of the electrons, plot the curve, and decide whether or not it cuts the axis of abscissae at a sharp angle. This is in principle the same method as is used in determining whether a given X-ray frequency appears suddenly at a given value of the energy of the electrons bombarding a solid; the curves of Fig. 4 were so obtained. Attempting to apply this same method to such a radiation as 2,536 of mercury, one has the solitary advantage that the frequency of the light is sharp and definite (it is not necessary to cut an arbitrary band of radiations out of a continuous spectrum) and two great counteracting disadvantages: the intensity of the light cannot be measured accurately (one has to guess it from the effect upon a photographic plate) and the impinging electrons never all have the same energy. Owing probably to these two difficulties, there is no published curve (that I know of) which cuts down across the axis of abscissae with such a decisive trend as the curves of Figs. 2 and 4. Still it is generally accepted that the advent of the single line is really sudden. The common argument is, that one can detect it on a photographic film exposed for a few hours when the energy of the bombarding electrons is (say) 5 equivalent volts, and not at all on a plate exposed for hundreds of hours when the bombarding voltage is (say) 4.5 volts. In this manner the energy of the electrons just sufficient to excite 2536 of mercury has been located at 4.9 equivalent volts. Dividing this critical energy (expressed in ergs) by the frequency of the radiation, we get

$$(4.9e/300) \,/\, (c/.00002536) = 6.59 \cdot 10^{-27} \qquad (5)$$

It agrees with the values of the constant which I designated by h in the two prior cases, and the data obtained with other kinds of atoms are not discordant. Gerlach arrives at $6.56 \cdot 10^{-27}$ as the mean of all values from experiments of this type upon many vapours. The evidence is not quite so strong as in the prior cases, but fortunately it is supplemented and strengthened by testimony of a new kind.

When electrons strike solids and excite X-rays, it is impossible to follow their own later history, or the adventures of a beam of radiation after it sinks into a metal. We have inferred that the electrons which collide with a piece of tungsten and disappear into it transfer their energy to X-rays, but the inference lacked the final support which would have been afforded by a demonstration of these very electrons, still personally present after the collision but deprived of their energy. Now when electrons are fired against mercury atoms, this demonstration is possible, and the results are very gratifying. I have already several times had occasion to remark, in this series of articles, that when an electron strikes a free atom of mercury, the result of the encounter is very different, according as its energy of motion was initially less than some 4.9 equivalent volts, or greater. In the former case, it rebounds as from an elastic wall, having lost only a very minute fraction of its energy, and this fraction spent in communicating motion to the atom; but in the latter case, it may and often does lose 4.9 equivalent volts of its energy *en bloc*, in a single piece as it were, retaining only the excess of its original energy over and above this amount. Thus if electrons of an energy of 4.8 equivalent volts are shot into a thin stratum of mercury vapor, nothing but electrons of that energy arrives at the far side; but if electrons of an only slightly greater energy, say 5.0 equivalent volts, are fired into the stratum, those which arrive at the far side will be a mixture of electrons of that energy, and very slow ones. The very slow ones can be detected by appropriate means, and the particular value of the energy of the bombarding electrons, at which some of them are for the first time transformed into these very slow ones, can be determined. Once more we meet that question as to whether the transformation does make its first appearance *suddenly*, but in this case the indications that it does are rather precise and easy to read. Furthermore it is possible to measure the energy of the slow electrons, and one finds that it is equal to the initial energy of the electrons, minus the amount 4.9 equivalent volts. (These measurements are not so exact as is desirable, and it is to be hoped that somebody will take up the task of perfecting them.)

We, therefore, see both aspects of the transaction which occurs when an electron whereof the energy is 4.9 equivalent volts, or greater, strikes a mercury atom. It loses 4.9 equivalent volts of energy, and we measure the loss; the atom sends forth radiation of a certain frequency, and no other; the atom does not send forth even this frequency of radiation, if none of the electrons fired against it has at least so much energy. We have already compared the energy transferred with the frequency radiated, and as in the case of X-rays excited from a solid target by very fast electrons, it is legitimate to say for these radiations which form the single-line spectra of metallic atoms, that

Excitation of the ray forming a single-line spectrum, by the collision of an electron against an atom, occurs as if *the energy in the radiation were concentrated in units of amount $h\nu$, and one such unit were created out of the total energy which the electron surrenders.*

There are yet several phenomena which I might treat by the same inductive method, arriving after each exposition at a Rule which would resemble one or the other of those which I have thus far written in italics; but it is no longer expedient, I think, to pass in each instance through the same elaborate inductive detour. These three phenomena which I have discussed already combine into an impressive and rather formidable obstacle to the classical manner of thinking. Here is a mercury atom, which receives a definite quantity of energy U from an electron, and distributes it in radiation of a definite frequency U/h. Here again is a multitude of atoms locked together into a solid, and when an electron conveys its energy U to the solid, it redistributes that energy in radiation of a definite frequency U/h. (It is true that many other radiations issue from the solid, but they are all explicable if one assumes that the electron may deliver over its energy in stages, and there is no radiation of the sort which would controvert the theory by virtue of its frequency exceeding U/h.) And when that radiation of frequency U/h in its turn strikes a metal, it is liable and able to release an electron from within the metal, conferring upon it an energy which is apparently equal to U. Apparently there is some correlation between an energy U and a frequency U/h, between a frequency ν and an energy $h\nu$. Apparently a block of energy of the amount U tends to pass into a radiation of the frequency U/h; apparently a radiation of the frequency ν tends to deliver up energy in blocks of the amount $h\nu$. The three italicized Rules coalesce into this one:

Photoelectric emission, and the excitation of X-rays from solids by electrons, and the excitation of single-line spectra from free atoms, occur as if *radiant energy of the frequency ν were concentrated into packets, or units, or corpuscles, of energy amounting to $h\nu$, and each packet were created in a single process and were absorbed in a single process.*

If the neutralizing *as if* were omitted, this would be the corpuscular theory *rediviva*. It is good policy to leave the *as if* in place for awhile yet. But conservatism such as this need not and should not deter anyone from using the idea as basis for every prediction that can be founded upon it, and testing every one of the predictions that can be tested by any possible way. Just so were the three phenomena cited in these Rules discovered. All of them involve either the emission or the absorption of radiation, and so do all the others which I could have quoted in addition, if this account had been written three years ago. Reserving to the end the one new phenomenon that transcends this limitation, I must explain the relation between this problem and the contemporary Theory of Atomic Structure.

The classical notion of a source of radiation is a vibrating electron. The classical conception of an atom competent to emit radiations of many frequencies is this: a family or a system of electrons, each electron remaining in an equilibrium-position so long as the system is not disturbed, one or more of the electrons vibrating when the system is jarred or distorted. A system with these properties would have to contain other things than electrons, otherwise it would fly apart; it would have to contain other things than particles of positive and particles of negative electricity intermixed, otherwise it would collapse together. One would have to postulate some sort of a framework, some imaginary analogue to a skeleton of springs and rods and pivots, to hold the electrons together in an ensemble able to vibrate and not liable to coalesce or to explode. This would not be satisfying, for in making atom-models one wants to avoid the elaborate machinery and in particular the non-electrical components; it would be much more agreeable to build an atom out of positive and negative electricity associated with mass, omitting all masses or structures not electrified. Nevertheless, if anyone had succeeded in devising a framework having the same set of natural frequencies as (say) the hydrogen atom exhibits in its spectrum—if anyone expert in dynamics or acoustics had been able to demonstrate that some peculiar shape of drumhead or bell, if anyone versed in electricity had been able to show that some particular arrangement of condensers and induction-coils has such a series of natural vibrations as some one kind of atom displays—then, it is quite safe to say, that framework or that membrane or that circuit would today be either the accepted atom-model, or at least one of the chief candidates for acceptance. Nobody ever succeeded in doing this; it is the consensus of opinion today that the task is an impracticable one.[9]

[9] It is difficult to put this statement into a more precise form. Rayleigh was of the opinion that the hydrogen spectrum could not be regarded as the ensemble of natural frequencies of a mechanical system, because it is the general rule for such systems that the *second* power of the frequency conforms to simple algebraic formulae, while in the hydrogen spectrum it is the *first* power for which the algebraic expression is simple. He admitted, however, that it was possible to find "exceptional" mechanical systems for which the first power of the frequency is given by a simple formula; which goes far to vitiate the conclusion. Another aspect of the formula (6) for

This set of natural frequencies which baffled all the efforts to explain it, the set constituting the two simplest of all spectra (the spectrum of atomic hydrogen and the spectrum of ionized helium), is given by the formula

$$\nu = R\left(\frac{1}{m^2} - \frac{1}{n^2}\right) \tag{6}$$

the different lines being obtained by assigning different integral values to the parameters m and n; lines corresponding to values of m ranging from 1 to 5 inclusive, and to values of n ranging from 2 to 40 inclusive, have already been observed, and there is no reason to doubt that lines corresponding to much higher values of m and n actually are emitted, but are too faint to be detected with our apparatus. The constant R has one value for hydrogen, another almost exactly four times as great for ionized helium.

Here, then, is the problem in its simplest presentation: How can a model for a hydrogen atom be constructed, which shall emit rays of the frequencies given by the formula (6), only these and no others? The obvious answer "By constructing a mechanical framework having precisely these natural frequencies" is practically excluded; it seems infeasible. Something radically different must be done. The achievement of Niels Bohr consisted in doing a radically different thing, with such a degree of success that the extraordinary divergence of his ideas from all foregoing ones was all but universally condoned. I do not know how Bohr first approached his theory; but it will do no harm to pretend that the manner was this.

Look once more at the formula for the frequencies of the hydrogen spectrum. It expresses each frequency as a difference between two terms, and the algebraic form of each term is of an extreme sim-

---

the hydrogen spectrum is this, that it specifies infinitely many frequencies within finite intervals enclosing certain critical values, such as R, $4R$, $9R$, and so forth. Poincaré is said to have proved that the natural frequencies of an elastic medium with a rigid boundary cannot display this feature, so long as the displacements are governed by the familiar equation $d^2\phi/dt^2 = k^2\nabla^2\phi$. For a membrane this equation is tantamount to the statement that the restoring-force acting upon an element of the membrane is proportional to the curvature of the membrane at that element. Ritz was able to show that the natural frequencies of a square membrane would conform to the formula (6), *if* the restoring-force upon each element of the membrane, instead of being proportional to the curvature of the membrane at that element, depended in an exceedingly involved and artificial manner upon the curvature of the membrane elsewhere. He apologized abundantly for the extraordinary character of the properties with which he had been obliged to endow this membrane, in order to arrive at the desired formula; but his procedure might have proved unsuspectedly fruitful, if Bohr's interpretation had not supplanted it.

plicity. Multiply now each member of the formula by h, that same constant h which we have encountered three times in the course of this article; and reverse the signs of the terms.[10] The formula becomes

$$h\nu = (-h\ R/n^2) - (-h\ R/m^2) \qquad (7)$$

In the left-hand member there stands $h\nu$. The reader will have become more or less accustomed to the notion that, under certain conditions and circumstances of Nature, radiant energy of the frequency ν apparently goes about in packets or corpuscles of the amount $h\nu$; now and then, here and there, energy is absorbed from such radiation in such amounts, or energy is converted into such radiation in such amounts. *Suppose that this also happens when a hydrogen atom radiates*, whatever the cause which sets it to radiating. Then the left-hand member of the equation (5) represents the energy which the hydrogen atom radiates; so also does the right-hand member; but the right-hand member is obviously the difference between two terms; *these terms are respectively the energy of the atom before it begins to radiate, and the energy of the atom after it ceases from radiating.*

The problem of the hydrogen atom has now experienced a fundamental change. The proposal to make a mechanical framework, having the natural vibration-frequencies expressed by (6), has been laid aside. The new problem, or the new formulation of the old problem, is this: how can a model for a hydrogen atom be constructed, which shall be able to abide only in certain peculiar and distinctive states or shapes or configurations, in which various states the energy of the atom shall have the various values $-hR$, $-hR/4$, $-hR/9$, $-hR/16$, and so forth?

Bohr's own model has become one of the best-known and most-taught conceptions of the whole science of physics, in the twelve years of its public existence. He based it upon the conception, then rapidly gaining ground and now generally accepted, that the hydrogen atom is a microcosmic sun-and-planet system, a single electron revolving around a much more massive nucleus bearing an electric charge equal in magnitude and opposite in sign to its own. This is really a most unpromising conception, very ill adapted to the modification we need to make. We want an atom which shall be able to assume only those definite values of energy which were listed above: $-hR$, $-hR/4$, $-hR/9$ and the rest. Now the energy of this sun-and-planet atom depends on the orbit which the electron is describing.

[10] For the explanation of this rather confusing reversal, see my third article (page 278; or page 11 of the reprint).

If the energy may assume only those definite values, the electron may describe only certain definite orbits. But there is no obvious reason why the electron should not describe any of an infinity of other orbits, circular or elliptical. To consider only the circular orbits: if the atom may have no other values of energy than $-hR$, and $-hR/4$, and $-hR/9$, and the rest of the series, then it may not revolve in any other circular orbits than those of which the radii are $e^2/2hR$, and $e^2/2(hR/4)$, and $e^2/2(hR/9)$, and so forth; but why just these? What prevents it from revolving in a circular orbit of radius $e^2/2(hR/2)$, or any other value not in the series? And for that matter how can it revolve in a closed orbit at all, since according to the fundamental notions of the electromagnetic theory it must be radiating its energy as it revolves, and so must sink into the nucleus in a gradually narrowing spiral?

Bohr did not resolve these difficulties, and no one has ever resolved them except by ignoring them. The customary procedure is to select some common feature of these permitted orbits, and declare that it is this feature which makes these orbits permissible, and forbids the electron to follow any other. For example, there is the fact that the angular momentum of the electron in any one of the permitted circular orbits is an integer multiple of the constant quantity $h/2\pi$, h being the same constant as we have met hitherto, which is hardly an accidental coincidence. If one could only think of some plausible reason why an electron should want to revolve only in an orbit where it can have some integer multiple of $h/2\pi$ for its angular momentum, and should radiate no energy at all while so revolving, and should refuse to revolve in an orbit where it must have a fractional multiple of $h/2\pi$, the model would certainly be much fortified. Failing this it is necessary to put this assertion about the angular momentum as a downright assumption, in the hope that its value will be so great and its range of usefulness so widespread that it will commend itself as an ultimate basic principle such as no one thinks of questioning. So far this hope has not been thoroughly realized. On the one hand, Sommerfeld and W. Wilson did succeed in generalizing it into a somewhat wider form, and using it in this wider form they explained the fine structure of the lines of hydrogen and ionized helium, and Epstein explained the effect of an electric field upon these lines. These are truly astonishing successes, and no one, I think, can work through the details of these applications to the final triumphant comparisons of theory with experiment, and not experience an impression amounting almost or quite to conviction. Yet on the other hand this generalization does not account for the frequencies forming the spectra of other elements.[11] There is the spectrum of neutral helium, for example, and the spectrum of sodium, and the spectrum of mercury; in each of these there are series of lines, of which the frequencies are clearly best expressed each as the difference between a pair of terms, and these terms should be the energies of the atom before and after radiating. But we have not the shadow of an idea what the corresponding configurations of the atom are; it may be that the outermost electron has certain permissible orbits, but we do not know what these orbits are like nor what common feature they possess.

Is it then justifiable to write down a Rule such as this: *the frequencies of the rays which free atoms emit are such as to confirm the idea that radiant energy of the frequency ν is emitted in packets or corpuscles of the amount $h\nu$?* Very few men of science, I imagine, would hesitate to approve this. However one may fluctuate in his feelings about Bohr's model of the atom, there always remains that peculiar relation among the frequencies emitted by the hydrogen atom, which is so nearly copied by analogous relations in the spectra of other elements. When one has once looked at the general formula

$$h\nu = \left(-\frac{hR}{n^2}\right) - \left(-\frac{hR}{m^2}\right) \tag{7}$$

and has once interpreted the first term on the right as the energy of an atom before radiating, the second term on the right as the energy of the atom after radiating, and the quantity $h\nu$ as the amount of the packet of energy radiated, it is very difficult to admit that this way of thinking will ever be superseded; particularly when one remembers the auxiliary facts, such as that fact about the electrons transferring just 4.9 equivalent volts to the mercury atoms which they strike, no more and no less. Analyzing the mercury spectrum in the same way as the hydrogen spectrum was analyzed, we find the frequencies expressible as differences between terms; interpreting the terms as energy-values, we find that between the normal state of the mercury atom and the next adjacent state, there is a difference in energy of 4.9 equivalent volts, and between this and the next adjacent state there is a further difference of 1.8 volts. This then is the reason why an electron with less than 4.9 equivalent volts of

[11] The mathematical experts who have laboured over the theory of the helium atom (two electrons and a nucleus of charge $+2e$) seem to have convinced themselves that the features which distinguish the permitted orbits of the electrons in this atom, whatever they may be, are definitely not the same features as distinguish the permitted orbits of the electron in the hydrogen atom. This cannot be said with certainty for any other atom.

energy can communicate no energy at all to a mercury atom; and an electron with 5 or 6 equivalent volts of energy can transfer only 4.9 of them. It is conceivable that other conditions may be found to govern the orbits of the electrons, so that the atoms shall have only the prescribed energy-values and no others; it is even conceivable that the conception of electron-orbits may be discarded; but the interpretation of the terms in the formula (7) as energies will, in all human probability, be permanent.

The foregoing Rule is thus very strongly based; but let us nevertheless rephrase it in a somewhat milder form as follows: *The idea that radiant energy of frequency ν is emitted in packets of the amount $h\nu$, and the contemporary theory of atomic structure, between them give a attractive and appealing account of spectra in general, and a convincingly exact explanation of two spectra in particular.*

But what has happened meanwhile to the Vibrator, to the oscillating electron, to the postulated electrified particle of which the vibrations caused light-waves to spread out from around it like sound waves from a bell? It has disappeared from the picture; or rather, since the attempt to account for the frequencies of a spectrum as the natural frequencies of an elastic framework was abandoned, no one has tried to re-insert it. But there are some who will never be quite happy with any new conception, until the vibrator is established as a part of it.

Ionization, the total removal of an electron from an atom, affords another chance to see whether radiant energy behaves as though it could be absorbed only in complete packets of amount $h\nu$. That it requires a certain definite amount of energy to deprive an atom of its loosest electron, an amount characteristic of the atom, may now be regarded as an experimental result quite beyond question, and not requiring the support of any special theory. Thus, a free-flying electron may remove the loosest electron from a free mercury atom which it strikes, if its energy amounts to 10.4 equivalent volts, not less; or the loosest electron from a helium atom if its energy amounts to at least 24.6 equivalent volts. If radiant energy of frequency ν goes about in parcels of magnitude $h\nu$, the frequency of a parcel which amounts just exactly to 10.4 equivalent volts is $\nu_o = 2.53.10^{16}$, corresponding to a wave length of 1188A. Light of inferior frequency should be unable to ionize a mercury atom; light of just that frequency should just be able to ionize it; light of a higher frequency ν should be able to ionize the atom, and in addition confer upon the released electron an additional amount of kinetic energy equal to $h(\nu - \nu_o)$. The same could be said, with appropriate numerical changes, for every other kind of atom. Of all the phenomena which might serve to illuminate this difficult question of the relations between radiation and atoms, this is the one which has been least studied. The experimental material is scanty and dubious. There is no reason to suppose that light of a lower frequency than the one I have called ν_0 is able to ionize; but it is not clear whether perceptible ionization commences just at the frequency ν_0, although it has been observed at frequencies not far beyond. The energy of the released electrons has not been measured.

The removal of deep-lying electrons, the electrons lying close to the nuclei of massive atoms, is much better known; and the data confirm in the fullest manner the idea that radiant energy of the frequency ν is absorbed in units amounting to $h\nu$. When a beam of X-rays of a sufficiently high frequency is directed against a group of massive atoms, various streams of electrons emanate from the atoms, and the electrons of each stream have a certain characteristic speed. The kinetic energy of each electron of any particular stream is equal to $h\nu$, minus the amount of energy which must be spent in extracting the electron from its position in the atom; for this amount of energy is independently known, being the energy which a free-flying electron must possess in order to drive the bound electron out of the atom, which is measurable and has been separately measured. Here again I touch upon a subject which has been treated in an earlier article of this series—the second—and to prevent this article from stretching out to an intolerable length, I refrain from further repetition of what was written there. The analogy of this with the photoelectric effect will escape no reader. Here as there, we observe electrons released with an energy which is admittedly not $h\nu$, but $h\nu$ minus a constant; the idea that this constant represents energy which the electrons have already spent in escaping, in one case through the surface of the metal and in the other case from their positions within atoms, is fortified by independent measurements of these energies which give values agreeing with these constants.

We have considered various items of evidence tending to show that radiant energy is born, so to speak, in units of the amount $h\nu$, and dies in units of the amount $h\nu$. Whether energy remains subdivided into these units during its incarnation as radiation remains unsettled; to settle this question absolutely, one would have to devise some way of testing the energy in a beam of radiation, otherwise than by absorbing it in matter; and such a way has not yet been discovered. There is, however, another quality which radiant energy possesses.

Conceive a stream of radiation in the form of an extremely long train of plane waves, flowing against a blackened plate facing normally against the direction in which they advance, which utterly absorbs them. This wave-train shall have an intensity I; by which it is meant, that an amount of energy I appears, in the form of heat, in unit area of the blackened plate in unit time. Furthermore, the radiation is found to exert a pressure p against the blackened plate; by which it is meant, that unit area of the plate (or the framework upholding it) acquires in unit time an amount of momentum p. According to the classical electromagnetic theory, verified by experience, p is equal to I/c. Unit area of the plate acquires, in unit time, energy to the amount I and momentum to the amount I/c.

Where is this energy, and where is this momentum, an instant before they appear in the plate? One might say that they did not exist, that they had vanished at the moment when the radiation left its source, not to reappear until it arrived at the plate; but such an answer would be contrary to the spirit of the electromagnetic theory, and we have long been accustomed to think of the energy as existing in the radiation, from the moment of its departure from the source to the moment of its arrival at the receiver; the term "radiant energy" implies this. Momentum has the same right to be conceived as existing in the radiation, during all the period of its passage from source to receiver. In the system of equations of the classical electromagnetic theory, the expression for the stream of energy through the electromagnetic field stands side by side with the expression for the stream of momentum flowing through the field. If the second expression is not so familiar as the first, and the phrase "radiant momentum" has not entered into the language of physics together with "radiant energy," the reason can only be that the pressure which light exerts upon a substance is very much less conspicuous than the heat which it communicates, and seems correspondingly less important,—which is no valid reason at all. Radiant energy and radiant momentum deserve the same standing; it is admitted that the energy I is the energy which is brought by the radiation in unit time to unit area of the plate which blocks the wave-train, and with it the radiation brings momentum I/c in unit time to unit area of the plate. The density of radiant energy in the wave-train is obviously I/c, the density of radiant momentum is I/c^2.[1]

Now let that tentative idea, that radiant energy of the frequency ν is emitted and absorbed in packets of the amount $h\nu$, be completed by the idea that these packets travel as entities from the place of their birth to the place of their death. Let me now introduce the word "quantum" to replace the alternative words *packet*, or *unit*, or *corpuscle*; I have held to these alternative words quite long enough, I think, to bring out all of their connotations. Then the energy I is brought to unit area of the plate, in unit time, by $I/h\nu$ of the quanta; which also bring momentum amounting to I/c. Shall we not divide up the momentum equally among the quanta as the energy is divided, and say that *each is endowed with the inherent energy* $h\nu$ *and with the inherent momentum* $h\nu/c$?

The idea is a fascinating one, but not so easy to put to the trial as one might at first imagine. None of the phenomena I have described in the foregoing pages affords any means of testing it. In studying the photoelectric effect, we concluded that each of the electrons released from an illuminated sodium plate had received the entire energy of a packet of radiation; but this does not imply that each of them had received the momentum associated with that energy; the momentum passed to the plate, to the framework supporting it, eventually to the earth. The same statement holds true for the release of electrons from the deep levels of heavy atoms, such as de Broglie and Ellis observed. Even if the same experiments should be performed on free atoms, as for example on mercury vapor, no clear information could be expected; for the momentum of the absorbed radiation may divide itself between the released electron and the residuum of the atom, and this last is so massive that the speed it would thus acquire is too low to be noticed. Only one way seems to be open; this is, to bring about an encounter between a quantum of radiation and a free electron, so that whatever momentum and whatever energy are transferred to the electron must remain with it, and cannot be passed along to more massive objects where the momentum, so far as the possibility of observing it goes, is lost. *A priori* one could not be certain that even this way is open; radiation might ignore electrons which are not tightly bound to atoms.

Arthur H. Compton, then of Washington University, is the physicist whose experiments were the first that clearly and strikingly disclosed such encounters between quanta of radiation and sensibly free electrons. Others had observed the effect which reveals them, but his were the first measurements accurate enough for inference. Unaware at the moment of the meaning of his data, he realized it almost immediately afterward, and so established the fact and the explanation both—a twofold achievement of a very unusual magnitude, whence the phenomenon received the name of "Compton effect" by a universal acceptance, and deservedly.

What Compton observed was not the presence of electrons possessed of momentum acquired from radiation—these electrons were however to be discovered later, as I shall presently mention—but the presence of radiation of a new sort, come into being by virtue of the encounters between the original radiation and free electrons. We have not encountered anything of this sort heretofore. When a quantum of radiant energy releases an electron from an atom, it dies completely and confers its entire energy upon the electron. The disposal of its momentum gives no trouble, for as I have mentioned the atom takes care of that. When the electron is initially free, and there is no atom to swallow up the momentum of the radiation, it cannot be ignored in this simple fashion. For if the quantum did utterly disappear in an encounter with a free electron, the velocity which the electron acquired would have to be such that its kinetic energy and its momentum were separately equal to the energy and momentum of the quantum; but these distinct two conditions would generally be impossible for the electron to fulfil. Hence in general, a quantum possessed of momentum cannot disappear by the process of transferring its energy to a free electron, whatever may be the case with an electron bound to a massive atom. This reflection might easily have led to the conclusion that radiation and free electrons can have nothing to do one with the other.

What actually happens is this: the energy and the momentum of the quantum are partly conferred upon the electron, the residues of each go to form a new quantum, of lesser energy and of lesser and differently-directed momentum, hence lower in frequency and deflected obliquely from the direction in which the original quantum was moving. The encounter occurs much like an impact between two elastic balls; what prevents the analogy from being perfect is, that when a moving elastic ball strikes a stationary one, it loses some of its speed but remains the same ball, whereas the quantum retains its speed but changes over into a new and smaller size. It is as though a billiard-ball lost some of its weight when it touched another but rolled off sidewise with its original speed. I do not know what this innovation would do to the technique of billiards, but it would at all events not make technique impossible; the result of an impact would still be calculable, though the calculations would lead to a new result. The rules of this microcosmic billiard-game in which the struck balls are electrons and the striking balls are quanta of radiant energy are definite enough to control the consequences. The rules are these:

Conservation of energy requires that the energy of the impinging quantum, $h\nu$, be equal to the sum of the energy of the resulting quantum, $h\nu'$, and the kinetic energy K of the recoiling electron. For this last quantity the expression prescribed by the special relativity-theory [12] is used, viz.

$$K = mc^2\left(\frac{1}{\sqrt{1-\beta^2}} - 1\right)$$

in which m stands for the mass of the electron and $c\beta = v$ for its speed. The equation of conservation of energy is then

$$h\nu = h\nu' + mc^2\left(\frac{1}{\sqrt{1-\beta^2}} - 1\right). \qquad (8a)$$

Conservation of momentum requires that the momentum of the impinging quantum be equal to the sum of the momenta of the resulting quantum and the recoiling electron. Momentum being a vector quantity, this rule requires three scalar equations to express it, which three may be reduced to two if we choose the x-axis to coincide with the direction in which the impinging quantum travels, and the y-axis to lie in the plane common to the paths of the recoiling electron and the resulting quantum. Designate by ϕ the angle between the paths of the impinging quantum and the recoiling electron; by θ the angle between the paths of the two quanta. The magnitude of the momentum-vector is, by the special relativity-theory, $mv/\sqrt{1-\beta^2}$. Conservation of momentum then requires:

$$\begin{aligned} h\nu/c &= (h\nu'/c)\cos\theta + \frac{mv}{\sqrt{1-\beta^2}}\cos\phi, \\ 0 &= (h\nu'/c)\sin\theta + \frac{mv}{\sqrt{1-\beta^2}}\sin\phi. \end{aligned} \qquad (8b)$$

Eliminating ϕ and v between these three equations, we arrive at this relation between ν and ν', the frequencies of the impinging quantum and the recoiling quantum—or, as I shall hereafter say, between the frequencies of the primary X-ray and the scattered X-ray—and the angle θ between the directions of the primary X-ray and the scattered X-ray:

$$\frac{\nu'}{\nu} = \frac{1}{1 + \frac{h\nu}{mc^2}(1-\cos\theta)}. \qquad (9)$$

[12] If the reader prefers to use the familiar expressions $\frac{1}{2}mv^2$ for the kinetic energy and mv for the magnitude of the momentum of the electron, he will arrive at a formula for ν' which, while apparently dissimilar to (9) and not so elegant, is approximately identical with it when v is not too large—or, which comes practically to the same thing, when $h\nu$ is small in comparison with mc^2; a condition which is realized for all X-rays now being produced

The relation between λ' and λ, the wavelengths of the primary beam and of the scattered beam, is still simpler, being

$$\lambda' - \lambda = \frac{h}{mc}(1 - \cos\theta). \tag{10}$$

The intrusion of this angle θ into the final equation may seem to contradict my earlier statement that the results of the impact are calculable; for it is true that there are not equations enough to eliminate θ, and yet I have offered no additional means of calculating it. In fact it cannot be calculated with the data at our command. All that we are able to say is that *if* the resulting quantum goes off in the direction θ, then its frequency is given by (9). What determines θ in any particular case? Reverting to the image of the billiard-balls, it is easy to see that the direction in which the rebounding ball rolls away depends on whether it gave a central blow, or a glancing blow, or something in between, to the initially stationary ball. If we knew just which sort of a blow was going to be given, we could calculate θ; otherwise we can only apply our conditions of conservation of energy and conservation of momentum to ascertain just how much of its energy the rebounding ball retains when θ has some particular value, and then produce—or, if we cannot produce at will, await—a collision which results in that value, and make our comparison of experiment with theory. So it is in this case of the rebounding quantum. When a beam of primary electrons is scattered by encountering a piece of matter, some quanta rebound in each direction, and all the values of θ are represented. We cannot know what determines the particular value of θ in any case; but we can at least select any direction we desire, measure the frequency of the quanta which have rebounded in that direction, and compare it with the formula. Fig. 6 is a diagram illustrating these relations.[13]

The comparison, which has now been made repeatedly by Compton, repeatedly by P. A. Ross, and once or oftener by each of several other physicists—notably de Broglie in Paris—is highly gratifying. The value of the frequency-difference between the primary X-rays and the scattered X-rays, that is to say, between the impinging quanta and the rebounding quanta, is in excellent accord with the formula, whether the measurements be made on the quanta recoiling at 45°, at 90° or at 135°, or at intermediate values of the angle θ. The method consists in receiving the beam of scattered X-rays into an X-ray spectroscope, whereby it is deflected against an ionization-chamber or a photographic plate at a particular point, of which the location is the measure of the wave-length. An image can be made on the same plate at the point where the beam would have struck it, if it had retained the frequency of the primary beam. The two images then stand sharply and widely apart. Indeed it is not necessary to make a special image to mark the place on the plate where a scattered beam of unmodified wave-length would fall, for there

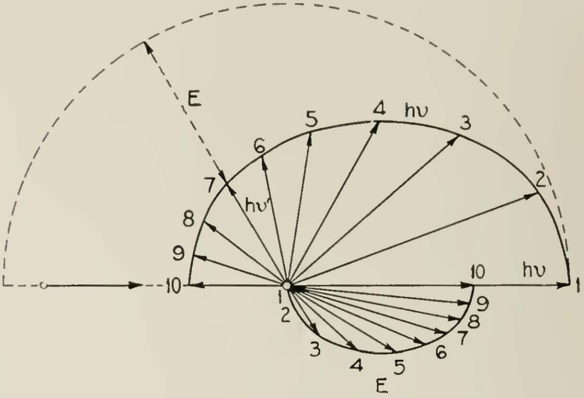


Fig. 6—Diagram showing the energy-relations ensuing upon an impact between a quantum and a free electron. (After Debye.) See footnote 13

nearly always is such a beam and such an image. A plausible explanation is easy to find; one has only to assume that the quanta composing this beam have rebounded from electrons so rigidly bound into atoms that they did not budge when the impinging quanta struck them, and these were reflected as from an immovable wall.[14]

[13] The diagram in Fig. 6 is designed to illustrate the relations between the energy of the primary quantum (radius of the dotted semicircle), the energy of the rebounding quantum (radius of the upper continuous curve), and the energy of the recoiling electron (radius of the lower continuous curve). Thus the two arrows marked with a 5 are proportional respectively to the energies of the secondary quantum and of the recoiling electron, when the encounter has taken place in such a fashion that the angle θ is equal to the angle between the arrow 10 and the upper arrow 5. In the same case, the angle between arrow 10 and lower arrow 5 is equal to ϕ of the equations (9).

[14] As a matter of fact we have no independent means of knowing that the recoiling electrons are initially free, or that the scattered beam with the modified frequency originates from collisions of primary quanta with initially free electrons; we know only that the frequency of the scattered quanta is such as would be expected if little or no energy is spent in freeing the electrons, and little or no momentum is transferred otherwise than to the electrons—which, of course, is not quite the same

In the photographs which I reproduce,[15] the imprints of these two beams stand side by side. In the first of them, Fig. 7, the spectrum of the primary rays is specially depicted on the upper half of the plate; one sees the α, β, and γ lines of the K-series of molybdenum, three lines (the first a doublet) of which the wavelengths are respec-

Fig. 7—Above, the K-spectrum of molybdenum (α-doublet, β-line, γ-line from left to right); below, the spectrum of this same radiation after scattering at 90° from aluminum (each line doubled). (P. A. Ross)

tively .710 − .711A, .633A, .618A. Below, the spectrum of the secondary rays scattered at the angle θ is spread out; to each of the primary rays there corresponds a scattered ray of the same wavelength, and beside it another ray of which the wavelength exceeds that of its companion by the required amount.

thing. The Compton effect has been demonstrated only where there are electrons associated with atoms. It may be that the rebound occurs only from an electron which is connected to an atom by some peculiar liaison, weak so far as the energy required to break it is concerned, but able to control the response of the electron to an impact. Something of this sort may have to be assumed to explain why the effect is apparently not greater for conductive substances than for insulating ones and is certainly feebler for massive atoms with numerous loosely-bound electrons than for light atoms with few.

[15] I am indebted to Professor Ross for these photographs.

Another series of photographs, in Fig. 8, shows the two scattered rays produced when a beam of the $K\alpha$-radiation of molybdenum falls upon various scattering substances: carbon (the sixth element of the periodic table), aluminium (the thirteenth), copper (the twenty-ninth), and silver (the forty-seventh). The relative intensity of the two rays—that is to say, the proportion between the number of quanta which rebound as from free electrons, and the number of quanta which recoil as from immobile obstacles—varies in a curious manner

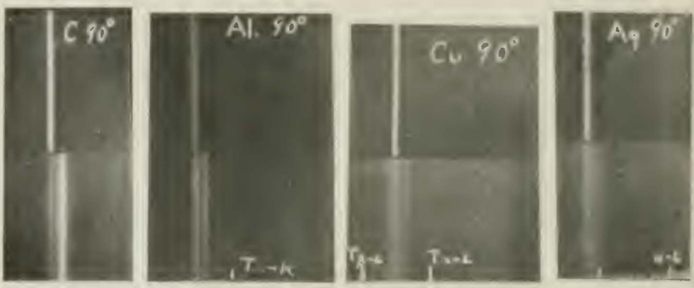


Fig. 8—Above, the K α-line of molybdenum; below, the same radiation after scattering at 90° from carbon, aluminium, copper and silver. (P. A. Ross)

from one of these elements to another. Most of the quanta scattered by lithium undergo the alteration in wavelength which we have calculated; nearly all of the quanta scattered by lead emerge with the same frequency as the incident quanta. Apparently, the heavier the atoms of a substance are, the less conspicuous does Compton's effect become. Further, the relative intensity of the two rays assumes different values for one and the same substance, depending on the direction of scattering. This is illustrated in Fig. 9, the curves of which may be interpreted as graphical representations of photographs like those of the foregoing Figure, the ordinate standing for the density of the image on the photographic plate. (Actually, the ordinate stands for a quantity which is much more nearly proportional to the true intensity of the rays—that is, the amount of ionization which they produce in a dense gas.) These curves show, in the first place, that the separation between the two scattered rays has the proper theoretical values at the angle 45°, at 90°, and at 135°; in the second place, among the quanta scattered at 45°, those that retain the primary wavelength are more abundant than the altered quanta, while among the quanta scattered at 135° the modified ones have the predominance. Why the relative commonness of these two kinds of scattering, of these two modes of interaction between quanta

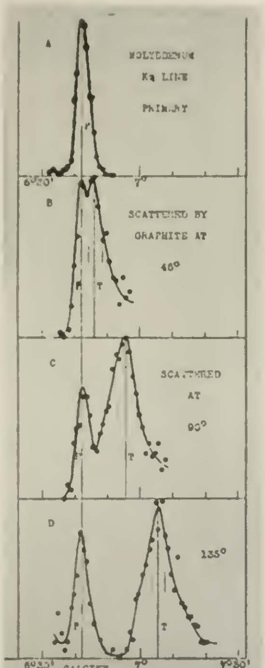


Fig. 4

Fig. 9 The modified and unmodified scattered rays, at various inclinations, recorded by the ionization-chamber method. The vertical line T represents the position calculated from (9) for the modified ray. (A. H. Compton, *Physical Review*)

and matter, should depend on the substance and on the angle θ is a deeper question than any we have considered.

The recoiling electrons also have been detected; and Figs. 10 and 11, which are photographs of the trails left by flying electrons as they

Fig. 10—Trails of recoiling electrons, mingled with long sinuous trails of electrons ejected from atoms by totally-absorbed quanta. (C. T. R. Wilson, *Proceedings of the Royal Society*)

proceed through air supersaturated with water vapor, shows evidence for these.[16] The long sinuous trails are those of fast electrons, which were liberated from their atoms by high-frequency quanta proceeding across the gas; each of these electrons possesses the entire energy of a vanished quantum (minus such part of it as was sacrificed when

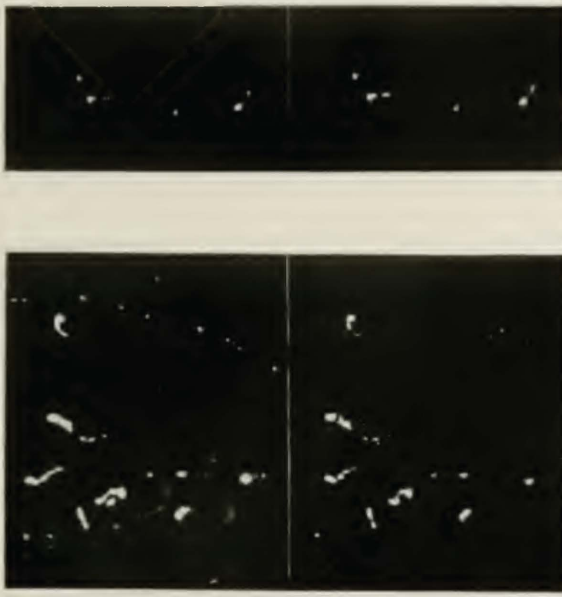


Fig. 11—Trails of recoiling electrons (C. T. R. Wilson, *Proceedings of the Royal Society*)

the electron emerged from its atom). The small slightly-elongated comma-like "blobs", the "fish tracks" as C. T. R. Wilson called them, are the trails of very slow electrons—these are the electrons from which quanta rebounded, transferring in the rebound a little of their energy and a little of their momentum. These appear only when the frequency of the X-ray quanta exceeds a certain minimum amount—a circumstance which, combined with others, shows that the com-

[16] I am indebted to Professor C. T. R. Wilson and to the Secretary of the Royal Society for permission to reproduce these photographs.

monness of the Compton effect depends not merely on the nature of the atoms and on the angle at which the scattering is observed, but also upon the frequency of the radiation. High-frequency quanta are liable to rebound in the manner prescribed by Compton's assumptions, but low-frequency quanta are not. Light of the visible spectrum suffers no change in wavelength when it is scattered.

Must we now concede that radiant energy travels about through space in the form of atom-like units, of corpuscles, of *quanta* every one of which, for a radiation of a specific frequency ν, possesses always the same energy $h\nu$ and always the same momentum $h\nu/c$? How indeed can we longer avoid admitting it? The phenomena which I have cited do certainly seem to close the case beyond any possibility of reopening it. Yet they might be interpreted in another way—a way which will probably seem extremely elaborate and artificial to the reader, a way which will seem like a mere excuse to avoid a simple and satisfying explanation; and yet this would not be sufficient to condemn it utterly. We might lay the whole blame and burden for all these "quantum" phenomena upon the atom. We might say that there is some mysterious mechanism inside every atom, which constrains it never to emit radiation of a frequency ν unless it has a quantity of energy $h\nu$ all packed up and ready to deliver, and never to absorb radiation of a frequency ν unless it has a special storeroom ready to receive just exactly the quantity of energy $h\nu$. This indeed is not a bad formulation of Bohr's theory of the atom. It would be necessary to go much further, and to say that not only every atom, but likewise every assemblage of atoms forming a liquid or a solid body, contains such a mechanism of its own; for the phenomena which I have called the "photoelectric effect" and the "inverse photoelectric effect" are qualities not of individual atoms, but of pieces of solid metal.[17] And it would be necessary to go much further yet, and make mechanisms to account for the transfer of momentum from radiation to electrons.

Yet even this would not be sufficient; for the most surprising and inexplicable fact of all is still to be presented. Here is the crux of the great dilemma. Imagine radiation of the frequency ν emerging from an atom, for a length of time determined by the condition that

[17] It was formerly contended that this explanation, while applicable to the behavior of free atoms which respond only to certain discrete frequencies, would not avail for a solid substance like sodium which delivers up electrons with energy $h\nu$, whatever the frequency ν may be. This contention, however, is probably not forcible, as it can be supposed that the solid has a very great number of natural frequencies very close together. This in fact was the inference from Epstein's theory of the photoelectric effect.

the total energy radiated shall be $h\nu$ exactly. According to the wave-theory, it emerges as a spherical wave-train, of which the wave-fronts are a series of expanding spheres, widening in all directions away from the atom at their common centre. Place another atom of the same kind some little distance away. Apparently it can absorb no radiant energy at all, unless it absorbs the whole amount $h\nu$ radiated from the first atom. But how can it do this, seeing that only a very small portion of each wavefront touched it or came any-where near it, and much of the radiant energy went off from the first atom in a diametrically opposite direction? How can it reach and suck up all the energy from the entire wavefront, so little of which it actually intercepts? And the difficulty with the momentum is even greater.

But, of course, this experiment is unrealizable. In any laboratory experiment, there are always great multitudes of radiating atoms close together, and the atoms exposed to the radiation are bathed in myriads of wave-trains proceeding from myriads of sources. Does then the atom which absorbs the amount $h\nu$ of energy take it in little bits, one from this wavetrain and another from that, until the proper capital is laid up? But if so, it surely would require some appreciable time to gather up the separate amounts. According to the classical electromagnetic theory, a bound electron placed in a wavetrain of wavelength λ will gather up energy from an area of each wavefront, of the order of magnitude of the quantity λ^2. Hence we should not expect that the exposed atom would finish the task of assembling the amount of energy $h\nu$ from the various wavetrains which pass by it, until the lapse of a time-interval sufficient for so much energy to flow against a circle of the area λ^2, set up facing the rays at the point where the atom stands. Set up a mercury arc, or better still, an X-ray tube, and measure the intensity of the radiation from it at various distances. You will easily find a position sufficiently near to it for convenience, and yet sufficiently far from it, so that if a circular target of this area were set in that position, the radiant energy falling upon it would not mount up in one minute—nor in one day—nor in one year, to the amount $h\nu$. Yet cover the source of rays with a shutter, and then put a piece of matter in that position, and then lift the shutter; and you will not have to wait a year, nor a day, nor a minute, for the first electron which emerges from the matter with a whole quantum of energy; it will come out so quickly that no experimenter has, as yet, demonstrated a delay. What possible assumptions about the structure of the *atom* can account for this?

More and more the evidence is piled up to compel us to concede that radiation travels around the world in corpuscles of energy $h\nu$ and momentum $h\nu/c$, which never expand, or at all events always remain small enough to be swallowed up in one gulp by an atom, or to strike an electron with one single concentrated blow.

But it is unfair to close the case without pleading once more the cause of the undulatory theory—the more so because, in the usual fashion, I have understated the old and presumptively familiar arguments in its favor, and given all the advantages to the arguments of the opposition, which still have the force and charm of novelty. Furthermore, I may have produced the impression that the conception of the quantum actually unites the corpuscular theory with the wave-theory, mitigating discord instead of creating it. Why are we not really voicing a perfectly competent wave-theory of light, when we imagine wave-trains limited both in length and in breadth, so narrow that they can dive into an atom, but so long that they contain $h\nu$ of energy altogether? *filamentary* wave-trains, so to speak, like the tracing of a sine-wave in chalk upon a blackboard, or the familiar picture of a sea-serpent?

Well, the difficulty is that the phenomena of interference and of diffraction, which are the basis of the wave-theory, imply that the wave-trains are broad, that they have a considerable cross-sectional area; these phenomena should not occur, if the wave-trains were filaments no thicker than an atom, or even so wide that their cross-sectional area amounted to λ^2. Let me cite one or two of these phenomena, in tardy justice to the undulatory theory, as a sort of a makeweight to all the "quantum" phenomena I have described. Imagine an opaque screen with a slit in it; light flows against the screen from behind, some passes through the slit. The slit may be supposed to be half a millimetre wide, or even wider. If light consists of quanta only as thick as an atom, or even as thick as the wave-length of the light, they will shoot through the slit like raindrops or sand-grains through a wide open skylight. If they are all moving in parallel directions before they reach the slit, they will continue so to move after they pass through it—for how shall they know that the slit has any boundaries, since they are so small and the slit is so large? The beam of light which has passed through the slit will always retain the same cross-section as the slit. But we know that in truth the beam widens after it goes through the slit, and it develops a peculiar distribution of intensity which is accurately the same as we should expect, if the wavefront is *wider* than the slit—so much wider, that the slit cuts a piece out of it, which piece spreads outwards independently in its own fashion.[18]. Therefore the quantum must be wider than the widest slit which displays clear diffraction-phenomena—and this makes it at least a millimetre wide! But this is not the limit! Cut another slit in the screen, parallel to the first one, a distance *d* away from it. Where the widening diffracted light-beams from the two slits interpenetrate one another, they will produce interference-patterns of light and shade, accurately the same as we should expect if the wavefront is wider than the distance *d*. The quantum must therefore be wider than the greatest distance between two slits, the light-beams passing through which are able to interfere with one another. The slits may be put quite far apart, and the light-beams brought together by systems of prisms and mirrors. This is the principle of Michelson's famous method of determining the diameters of stars. He obtained interference fringes when the two beams of light were taken from portions of the wavefront *twenty feet apart!*[19]

Therefore the quantum is twenty feet wide! This is the object from which an atom one ten-millionth of a millimetre wide can suck up all its energy! this is what enters as a unit into collision with an electron ten thousandfold smaller yet!

The evidence is now before the reader; not the entire evidence for either of the two conceptions of radiation, but, I think, a fair sampling for both. If either view has been inequitably treated, it is the undulatory theory which has been underrated; for, as I have said already but cannot say too often, the evidence that light partakes of the nature of a wave-motion is tremendously extensive and tremendously compelling; it seems the less powerful only because it is so thoroughly familiar, and through much repetition has lost the force of novelty. Still, it is not necessary to hold all the relevant facts continually in mind. If one could reconcile a single typical fact of the one sort, such as the interference between beams of light brought together from parallel courses far apart, with a single outstanding fact of the other sort, such as the instantaneous emergence of electrons with great energy from atoms upon which a feeble beam of light has only just been directed—if one could unify two such phenomena as these, all of the others would probably fuse spontaneously into a harmonious system. But in thinking about these things, there is one more all-important

[18] One might, of course, inquire, why should a *piece* of the wavefront of a quantum, cut out of it by the edges of a slit, expand after passing through the slit when the quantum itself apparently rushes through space without expanding?

[19] It might be argued that these quanta from stars have come an enormously long way, and possibly have had a better chance to expand than the quanta passing across a laboratory room from an X-ray tube or a mercury arc to a metal plate. However, since the photoelectric cell is used to measure the brightness of a star, they evidently produce the same sort of photoelectric effect as newborn quanta.

fact that must never be forgotten: the quantum-theory involves the wave-theory in its root and basis, for *the quantum of a given radiation is defined in terms of the frequency of that radiation, and the frequency* is determined from the wavelength, and the wavelength *is determined by applying the wave-theory* to measurements on interference and diffraction patterns. Was there ever an instance in which two such apparently contradictory theories were woven so intimately the one with the other!

The fusion of the theories is not likely to result from new experimental evidence. Indeed there are already indications that further experiments will merely accentuate the strangeness, much as happened with the numerous experiments devised and performed three or four decades ago in the hope of settling whether the earth does or does not move relatively to the aether. More probably what is required is a modification, indeed a revolutionary extension in the art of thinking such a revolution as took place among a few mathematicians when non-Euclidean geometry was established by the side of Euclidean, as is taking place today among the disciples of Einstein who are striving to unlearn the habitual distinctions between time and space—such a revolution, to go centuries back into the past, as occurred in the minds of men generally when they learned to realize that the earth is round, and yet at every place upon it the sky is above and the ground is below. Our descendants may think pityingly of us as we of our ancestors, who could not comprehend how a man can stand upright at the Antipodes.

Wave Propagation Over Parallel Tubular Conductors: The Alternating Current Resistance

By SALLIE PERO MEAD

SYNOPSIS: On the basis of Maxwell's laws and the conditions of continuity of electric and magnetic forces at the surfaces of the conductor, the fundamental equations are established for the axial electric force and the tangential magnetic force in a non-magnetic tubular conductor with parallel return. The alternating current resistance per unit length is then derived as the mean dissipation per unit length divided by the mean square current. The general formula is expressed as the product of the alternating current resistance of the conductor with concentric return and a factor, termed the "proximity effect correction factor," which formulates the effect of the proximity of the parallel return conductor. The auxiliary functions which appear in the general formula are each given by the product of the corresponding function for the case of a solid wire and a factor involving the variable inner boundary of the conductor.

In general, the resistance may be calculated from this formula, using tables of Bessel functions. The most important practical cases, however, usually involve only the limiting forms of the Bessel functions. Special formulae of this kind are given for the case of relatively large conductors, with high impressed frequencies, and for thin tubes. A set of curves illustrates the application of the formulae.

I. INTRODUCTION

WHERE circular conductors of relatively large diameter are under consideration, the effect on the alternating current resistance of the tubular as distinguished from the solid cylindrical form becomes of practical importance. Mr. Herbert B. Dwight has worked on a special case of this problem and developed a formula for the ratio of alternating to direct current resistance in a circuit composed of two parallel tubes when the tubes are thin.[1] As infinite sums of infinite series are involved, however, his result is not well adapted to computation.

Mr. John R. Carson has given a complete solution for the alternating current resistance of two parallel solid wires in his paper "Wave Propagation Over Parallel Wires: The Proximity Effect," *Phil. Mag.*, April, 1921. The analysis of that paper may readily be extended to the more general case of propagation over two tubular conductors by a parallel method of development. This is done in the present paper. As the underlying theory is identical in the two problems, familiarity with the former paper will be assumed and the analysis will merely be sketched after the fundamental equations are established.

[1] "Proximity Effect in Wires and Thin Tubes," Trans. A. I. E. E., Vol. XLII (1923), p. 850.

In this paper formulae for the alternating current resistance have been worked out in detail with particular reference to the case of relatively large conductors at high frequencies and to relatively thin tubes. In general the auxiliary functions involved are expressed as the product of the corresponding functions for solid wires by a correction factor which formulates the greater generality due to the variable inner boundary of the conductors. As far as possible the symbols are the same as in the solid wire case but refer now to the system of tubular conductors. Primes are added where the letters denote the corresponding functions for the solid wire case. This will hardly lead to confusion with the primes used in connection with the Bessel functions to denote differentiation.

The general solution is developed in section II. The alternating current resistance of one of the tubular conductors is expressed as the product of the alternating current resistance of the conductor with concentric return and a factor which formulates the effect of the proximity of the parallel return conductor. Section III is a summary of the general formula, special asymptotic forms and forms for thin conductors.

II. Mathematical Analysis and Derivation of Formulae

We require the expression for the axial electric force, E_z, in the conductors. Since the tubular conductor does not extend to $r=O$, the electric force must be expressed by the more general Fourier-Bessel expansion,

$$E_z = \sum_{n=O}^{\infty} A_n \left[J_n(\rho)+\lambda_n K_n(\rho)\right] \cos n\theta,$$

where

$$\begin{aligned}\rho &= ir\sqrt{4\pi\lambda\mu i\omega}\\ &= \xi = xi\sqrt{i} \text{ when } r=a\\ &= \zeta = yi\sqrt{i} \text{ when } r=\alpha,\end{aligned}$$

a and α being the outer and inner radii, respectively, of the conductors. The additional set of constants $\lambda_o, \lambda_1 \ldots \lambda_n$ is to be determined by the conditions of continuity at the inner boundary of the conductor. It is necessary to satisfy the boundary conditions at the surface of one conductor only, since the symmetry of the system insures that they will then be satisfied at the surface of the other also.

In the dielectric space inside the tube where $r<\alpha$, the axial electric force may be written

$$E_z=\sum_{n=0}^{\infty} C_n J_n(\rho)\cos n\theta, \tag{1}$$

or replacing the Bessel functions by their values for vanishingly small arguments,

$$E_z=\sum_{n=0}^{\infty} D_n r^n \cos n\theta \tag{2}$$

where $D_0, D_1 \ldots D_n$ are constants determined by the boundary conditions. Applying Maxwell's law relating the normal and tangential magnetic forces H_r and H_θ to the axial electric force, gives

$$\mu i\omega H_\theta=\frac{\rho}{r}\sum_{n=0}^{\infty} A_n\left[J_n'(\rho)+\lambda_n K'_n(\rho)\right]\cos n\theta, \tag{3}$$

$$\mu i\omega H_r=\frac{1}{r}\sum_{n=0}^{\infty} A_n\left[J_n(\rho)+\lambda_n K_n(\rho)\right]\sin n\theta, \tag{4}$$

for the space inside the conductor, and

$$i\omega H_\theta=\sum_{n=0}^{\infty} nD_n r^{n-1}\cos n\theta, \tag{5}$$

$$i\omega H_r=\sum_{n=0}^{\infty} nD_n r^{n-1}\sin n\theta, \tag{6}$$

for the inner dielectric ($\mu=1$). Equating the two expressions for the tangential magnetic force H_θ and for the normal magnetic induction μH_r term by term at the surface $r=\alpha$,

$$\left[\zeta J_n'(\zeta)-\mu n J_n(\zeta)\right]+\lambda_n\left[\zeta K_n'(\zeta)-\mu n K_n(\zeta)\right]=0. \tag{7}$$

Whence, for the practically important case of non-magnetic conductors in which $\mu=1$, we have

$$\lambda_n=-\frac{J_{n+1}(\zeta)}{K_{n+1}(\zeta)} \tag{8}$$

and

$$E_z=\sum_{n=0}^{\infty} A_n\left[J_n(\rho)-\frac{J_{n+1}(\zeta)}{K_{n+1}(\zeta)}K_n(\rho)\right]\cos n\theta. \tag{9}$$

In the subsequent analysis $J_n(\xi)$ of the solution for the solid wire case is replaced by

$$J_n(\xi) - \frac{J_{n+1}(\zeta)}{K_{n+1}(\zeta)} K_n(\xi) = M_n(\xi), \tag{10}$$

and $J_n'(\xi)$ is replaced by

$$J_n'(\xi) - \frac{J_{n+1}(\zeta)}{K_{n+1}(\zeta)} K_n'(\xi) = M_n'(\xi). \tag{11}$$

Otherwise the formulation of the alternating current resistance of the conductor proceeds exactly as in the solid wire case. For the electric force at the surface $r=a$ in the conductor, we write

$$E_z = A_o[M_o(\xi) + h_1 M_1(\xi)\cos\theta + h_2 M_2(\xi)\cos 2\theta + \ldots] \tag{12}$$

and determine the fundamental coefficient A_o in terms of the current in the conductor. The resistance R of the tubular conductor per unit length is defined as the mean dissipation per unit length divided by the mean square current where the mean dissipation is calculated by Poynting's theorem. Accordingly, we get

$$R = \text{Real}\frac{2\mu i\omega}{\xi}\left\{\frac{M_o(\xi)}{M_o'(\xi)} + \frac{1}{2}\sum_{n=1}^{\infty} |h_n|^2 \frac{M_n(\xi)}{M_o'(\xi)} \text{conj.} \frac{M_n'(\xi)}{M_o'(\xi)}\right\}. \tag{13}$$

To determine the harmonic coefficients $h_1 \ldots h_n$ or $A_1 \ldots A_n$, the total tangential magnetic force and the total normal magnetic induction at the outer surface of a conductor are expressed in terms of the coordinates of that conductor alone, and the conditions of continuity at the surface are applied. This leads to the set of equations

$$q_n = (-1)^n 2\rho_n k^n - \frac{(-1)^n}{(n-1)!}\rho_n k^n \sum_n (q) \qquad n = 1, 2, 3 \ldots \infty \tag{14}$$

where

$$\sum_n (q) = \frac{n!}{1!} k q_1 - \frac{(n+1)!}{2!} k^2 q_2 + \ldots,$$

$$\sigma_n = (\xi M_n'(\xi) - n\mu M_n(\xi)) / \xi M'(\xi),$$

$$\rho_n = (\xi M_n'(\xi) - n\mu M_n(\xi)) / (M_n'(\xi) + n\mu M_n(\xi)),$$

$$q_n = \sigma_n h_n,$$

$$\frac{a}{c} = k.$$

When the permeability is unity, the solution, to the same order of approximation as in the solid wire case, is

$$|h_n|^2 = \frac{u_1^2+v_1^2}{u_{n-1}^2+v_{n-1}^2}\left|\frac{1+\lambda_1 K_1(\xi)/J_1(\xi)}{1+\lambda_n K_{n-1}(\xi)/J_{n-1}(\xi)}\right|^2 p_n^2(1+2ngk^2/s^{n-1}) \qquad (16)$$

where

$$g = \frac{\sqrt{2}}{x}\,\frac{p[u_1(u_o+v_o)-v_1(u_o-v_o)]-q[u_1(u_o-v_o)+v_1(u_o+v_o)]}{u_o^2+v_o^2}, \qquad (17)$$

$$p+iq = \frac{1+\lambda_1 K_1(\xi)/J_1(\xi)}{1+\lambda_1 K_o(\xi)/J_o(\xi)}, \qquad (18)$$

$$J_n(\xi) = u_n + iv_n,$$

$$p_n = (-1)^n 2k^n s^n, \quad n = 1, 2 \ldots \infty,$$

$$s = 2\,\frac{1-\sqrt{1-(2k)^2}}{(2k)^2}.$$

Since the resistance R_o of an isolated tubular conductor is given by

$$R_o = \text{Real}\,\frac{2\mu i p}{\xi}\,\frac{M_o(\xi)}{M_o'(\xi)} \qquad (19)$$

equation (13) becomes equation (1) of the formulae in the next section. This is the general solution for the case of non-magnetic conductors.

In general R may be calculated from this formula and tables of Bessel functions. The ber, bei, ker and kei functions [2] and the recurrence formulae are sufficient to evaluate the Bessel functions but the process is long. In the most important practical cases, the conductors are rather large and the applied frequencies fairly high. When this is true as well as when the tubes are very thin the formulae usually involve only the limiting forms of the Bessel functions. These special results are given in the next section.

III. Alternating Current Resistance Formulae for Non-Magnetic Conductors

The symbols used are:

a = outer radius of conductor in centimeters,

α = inner radius of conductor in centimeters,

c = interaxial separation between conductors in centimeters,

$k = a/c$

λ = conductivity of conductor in electromagnetic c.g.s. units,

[2] A convenient table of these functions for arguments from 0 to 10 at intervals of 0.1 is incorporated in Mr. Dwight's paper "A Precise Method of Calculation of Skin Effect in Isolated Tubes," *J. A. I. E. E.*, Aug., 1923.

μ = permeability of conductor in electromagnetic c.g.s. units,

$\omega = 2\pi$ times frequency in cycles per second,

$i = \sqrt{-1}$

$x = a\sqrt{4\pi\lambda\omega}$

$y = \alpha\sqrt{4\pi\lambda\omega}$

$\xi = xi\sqrt{i}$

$\zeta = yi\sqrt{i}$

$\lambda_n = -J_{n+1}(\zeta)/K_{n+1}(\zeta)$

$J_n(\xi) = u_n + iv_n$

= Bessel function of first kind of order n and argument $xi\sqrt{i}$,

$J_n'(\xi) = \dfrac{dJ_n(\xi)}{d\xi}$

$u_n' + iv_n' = \dfrac{dJ_n(\xi)}{dx}$

$K_n(\xi)$ = Bessel function of second kind of order n and argument $xi\sqrt{i}$,

$K_n'(\xi) = \dfrac{dK_n(\xi)}{d\xi}$

R = resistance per unit length of tubular conductor with parallel return,

R_o = resistance per unit length of tubular conductor with concentric return in electromagnetic c.g.s. units,

C = proximity effect correction factor,

$$R = C\,R_o. \tag{1}$$

The auxiliary functions involved are:

$$^{3}\ R_o = R_o'm\left(1 - \frac{n}{m}\,\frac{u_o u_o' + v_o v_o'}{u_o v_o' - u_o' v_o}\right) \tag{20}$$

where

$$R_o' = \frac{1}{a}\sqrt{\frac{\omega}{\pi\lambda}}\,\frac{u_o v_o' - u_o' v_o}{u_1^2 + v_1^2} \tag{21}$$

= resistance of solid wire with concentric return,

$$m + in = \frac{1 + \lambda_o K_o(\xi)/J_o(\xi)}{1 + \lambda_o K_o'(\xi)/J_o'(\xi)}, \tag{22}$$

$$g = g'p\left\{1 - \frac{q\,[u_1(u_o - v_o) + v_1(u_o + v_o)]}{p\,[u_1(u_o + v_o) - v_1(u_o - v_o)]}\right\}, \tag{23}$$

[3] The ratio R_o/R_o' oscillates about unity which it approaches more and more closely as the frequency increases. It is due to the fact that the phase of the current in the inner portion of the solid conductor may be such as to oppose the current in the outer portion, that the resistance of the solid conductor may be greater than that of the tube even though the heating effect in the latter is the greater.

where

$$g' = \frac{\sqrt{2}}{x}\,\frac{u_1(u_o+v_o)-u_1(u_o-v_o)}{u_o^2+v_o^2}, \tag{24}$$

$$p+iq = \frac{1+\lambda_1 K_1(\xi)/J_1(\xi)}{1+\lambda_1 K_o(\xi)/J_o(\xi)}, \tag{25}$$

$$w_n = w_n' \frac{a_n}{[1+\lambda_n K_{n-1}(\xi)/J_{n-1}(\xi)]^2}\left(1-\frac{b_n}{a_n}\,\frac{u_n u_n'+v_n v_n'}{u_n v_n'-u_n' v_n}\right), \tag{26}$$

where

$$w_n' = \frac{u_n v_n' - u_n' v_n}{u_{n-1}^2+v_{n-1}^2}, \tag{27}$$

$$a_n + ib_n = \left(1+\lambda_n \frac{K_n(\xi)}{J_n(\xi)}\right)\text{conj.}\left(1+\lambda_n \frac{K_n'(\xi)}{J_n'(\xi)}\right), \tag{28}$$

$$s = 2\,\frac{1-\sqrt{1-(2k)^2}}{(2k)^2}. \tag{29}$$

The formula for the correction factor C is then

$$C = 1+\frac{2}{aR_o}\sqrt{\frac{\omega}{\pi\lambda}}(S_1+2gk^2S_2) \tag{II}$$

where

$$S_1 = \sum_{n=1}^{\infty} w_n k^{2n} s^{2n}, \tag{30}$$

$$S_2 = \sum_{n=1}^{\infty} n w_n k^{2n} s^{n+1}. \tag{31}$$

For large values of the argument

$$R_o = R_o'\left[m-n\left(1-\frac{1}{\sqrt{2}x}\right)\right] \tag{32}$$

and the correction factor is

$$C = 1+2\,\frac{\sqrt{2}-1/x}{m-n(1-1/\sqrt{2}x)}\left(S_1-\frac{2\sqrt{2}}{x}\left[p+q\left(1-\frac{1}{\sqrt{2}x}\right)\right]k^2S_2\right) \tag{III}$$

When x and y are both large quantities, the auxiliary functions are as follows, provided terms of the second order in $1/x$ and $1/y$ are negligible, n in d and h below being equal to the number of terms in which S_1 and S_2 converge to a required order of approximation.

With the notation

$$\cos = \cos\sqrt{2}(x-y),$$

$$\sin = \sin\sqrt{2}(x-y),$$

$$\exp = \exp[-\sqrt{2}(x-y)],$$

$$R_o = R_o'\,\frac{1+[(1+a)\sin-(1-a)\cos]\exp - a\exp^2}{1-[(1-b)\sin+(1+b)\cos]\exp + b\exp^2} \tag{33}$$

where

$$a = 1 - \frac{1}{2\sqrt{2x}} - \frac{3}{2\sqrt{2y}},$$

$$b = 1 + \frac{3}{2\sqrt{2x}} - \frac{3}{2\sqrt{2y}},$$

$$\frac{1}{aR_o'}\sqrt{\frac{\omega}{\pi\lambda}} = \sqrt{2} - \frac{1}{x}, \tag{34}$$

$$g = g' \frac{1 + [(1-c)\cos - (1+c)\sin]\exp - c\exp^2}{1 - [(1+c)\cos + (1-c)\sin]\exp + c\exp^2}, \tag{35}$$

where

$$c = 1 - \frac{1}{2\sqrt{2x}} - \frac{15}{2\sqrt{2y}},$$

$$g' = -\sqrt{2}/x, \tag{36}$$

$$w_n = w_n' \frac{1 - [(1-d)\cos - (1+d)\sin]\exp - d\exp^2}{1 - [(1+h)\cos + (1-h)\sin]\exp + h\exp^2}, \tag{37}$$

where

$$d = 1 + \frac{4n^2 - 1}{2\sqrt{2x}} - \frac{4(n+1)^2 - 1}{2\sqrt{2y}},$$

$$h = 1 + \frac{4(n-1)^2 - 1}{2\sqrt{2x}} - \frac{4(n+1)^2 - 1}{2\sqrt{2y}},$$

$$w_n' = \frac{1}{\sqrt{2}} - \frac{2n-1}{2x}. \tag{38}$$

At frequencies sufficiently high to afford practically skin conduction, the following formulae indicate the way in which the resistance of the tubular conductor approaches its limit, the resistance of the solid wire.

$$R_o = R_o' \frac{1 + 2\sin\exp}{1 - 2\cos\exp}, \tag{39}$$

$$\frac{1}{aR_o'}\sqrt{\frac{\omega}{\pi\lambda}} = \sqrt{2} - \frac{1}{x},$$

$$C = C_m(1 - A/x), \tag{IV}$$

$$C_m = \frac{1 + k^2s^2}{1 - k^2s^2}, \tag{40}$$

$$A = 2\sqrt{2}\,\frac{k^2s^2}{1 - k^4s^4}\left\{1 + 2k^2\frac{(1-k^2s^2)^2}{(1-k^2s)^2}\,\frac{1 - 2\sin\exp}{1 - 2\cos\exp}\right\}. \tag{41}$$

When the conductors are very thin tubes, i.e., thin as compared to the radius, $(a-\alpha)/a$ is necessarily small and, in general, $x-y$ is small. Of course, when the frequency is high enough, $x-y$ becomes large in any case. When this is true with respect to thin tubes, however, x and y will usually be large enough to make the asymptotic formulae applicable; but, if $x-y$ is small, the approximations

$$J_n(\zeta)=J_n(\xi)-(\xi-\zeta)J_n'(\xi)+\frac{(\xi-\zeta)^2}{2!}J_n''(\xi),$$

$$K_n(\zeta)=K_n(\xi)-(\xi-\zeta)K_n'(\xi),+\frac{(\xi-\zeta)^2}{2!}K_n''(\xi),$$

reduce the correction factor to

$$C=1+2\beta^2 f\left\{\sum_{n=1}^{\infty}k^{2n}s^{2n}\frac{d_n}{D_n}-2k^2\frac{\frac{4d_1^2}{x^4}-\beta^3c_1}{D_1}\sum_{n=1}^{\infty}k^{2n}s^{n+1}n\frac{d_n}{D_n}\right\} \qquad \text{(V)}$$

where $\beta=\dfrac{a-\alpha}{a}$,

$$f=\frac{(1+\beta/2)^2}{1+\beta+\beta^2}=\frac{c_o^2}{d_o},$$

$$D_n=\beta^2c_n^2+\frac{4n^2}{x^4}d_n^2,$$

$$c_n=1+\frac{2n+1}{2}\beta,$$

$$d_n=1+(n+1)\beta+\frac{(n+1)(n+2)}{2}\beta^2.$$

and the resistance with concentric return to

$$R_o=\frac{1}{2\pi\lambda a(a-\alpha)}\,\frac{1+\beta+\beta^2}{1+\beta/2}. \qquad (42)$$

$1/2\pi\lambda a(a-\alpha)$ is, of course, the direct current resistance of a very thin conductor.

If $(a-\alpha)/a$ is very small and negligible compared with $2n/x^2$, where n is the number of terms in which the series of (V) converge to a required order of approximation,

$$C=1+\frac{x^4}{2}\left(\frac{a-\alpha}{a}\right)^2\left\{\begin{array}{l}\left(1-\dfrac{a-\alpha}{a}\right)\left\{\displaystyle\sum_{n=1}^{\infty}\frac{k^{2n}s^{2n}}{n^2}+2k^2s\log(1-k^2s)\right\}\\[2ex] \quad+\dfrac{a-\alpha}{a}\left\{\log(1-k^2s^2)+2\dfrac{k^4s^2}{1-k^2s}\right\}\end{array}\right\} \qquad \text{(VI)}$$

As a check on formulae (V) and (VI), the limiting cases may be arrived at directly as follows. If the conductors are thin tubes, the harmonic coefficients are given by

$$h_n=(-1)^{n+1}2k^n\frac{\xi-\zeta}{\frac{2n}{\xi}-(\xi-\zeta)\left(1-\frac{2n(n+1)}{\xi^2}\right)}$$

$$-(-1)^{n+1}\frac{\xi-\zeta}{\frac{2n}{\xi}-(\xi-\zeta)\left(1-\frac{2n(n+1)}{\xi^2}\right)}k^n\left[nkh_1-\frac{n(n+1)}{2!}k^2h_2+\ldots\right]. \quad (43)$$

When ξ is very large

$$h_n=(-1)^n2k^n\left[1-\frac{1}{2}\left\{nkh_1-\frac{n(n+1)}{2!}k^2h_2+\ldots\right\}\right]$$

$$=(-1)^n2k^ns^n, \quad (44)$$

and

$$\frac{M_n}{M_o}=\frac{M_n'}{M_o'}=1 \quad (45)$$

so that

$$C=\text{Real}\left[1+\frac{1}{2}\sum_{n=1}^{\infty}|h_n|^2\frac{M_n}{M_o}\text{conj.}\frac{M_n'}{M_o'}\right]$$

$$=\frac{1+k^2s^2}{1-k^2s^2}, \quad (46)$$

the same result as for the corresponding limiting case of a solid conductor.

On the other hand, if ξ is not large and $\xi-\zeta$ is very small,

$$h_n=(-1)^{n+1}\frac{k^n}{n}\xi(\xi-\zeta), \quad (47)$$

$$\frac{M_n}{M_o}=1, \quad (48)$$

$$\frac{M_n'}{M_o'}=-\frac{in}{x(x-y)}, \quad (49)$$

so that

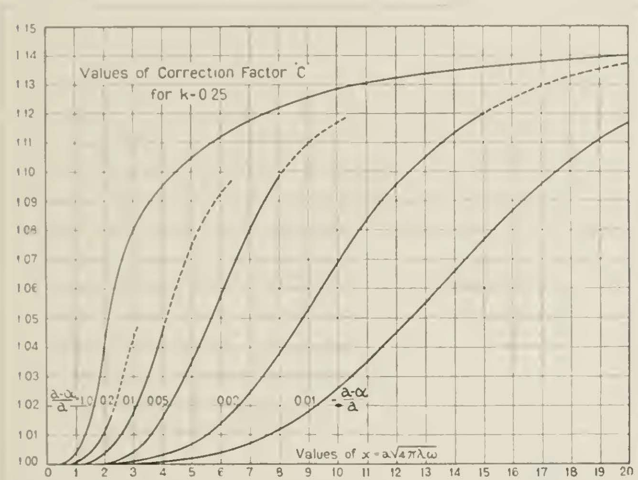
$$C=1, \quad (50)$$

and

$$R=R_o=R_{d.c.}, \quad (51)$$

where R_{dc} is the direct current resistance of the thin tubular conductor. Eqs. (46) and (50) agree with the corresponding limits of formulae V and VI respectively.

The curves of the accompanying figure do not pretend to represent the proximity effect correction factor with precision. They are, however, accurate for thin tubes, and indicate the order of magnitude of the factor for various values of the thickness of the tubular conductor and show the nature of its variation with respect to the applied



frequency. They are computed from formula (V) which is valid for quite high frequencies when the tubes are thin. When the thickness of the tubes is greater, however, the range of validity with respect to frequency is smaller, the dotted portions indicating a doubtful degree of precision. It was previously pointed out in connection with formula (IV) and is immediately deducible from physical considerations, that all of the curves eventually coincide with the curve for the solid wire which approaches the value 1.155 asymptotically.

As a simple application, suppose the resistance is required of a tubular conductor with an outer radius of 0.4125 cm. (that of No. 0 gauge A.W.G. copper wire) whose resistivity is 1696.5 electromagnetic units per cm., where there is an equal parallel return so situated that k = 0.25 and a frequency of 5,000 cycles per second is applied to the circuit. Then $m=\sqrt{4\pi\lambda\omega}=15.26$ and $x=ma=15.26\times0.4125=6.30$. When the ratio of the thickness of the conductor to the radius is greater than about 0.01 the proximity effect correction factor C is appreciable. If the ratio is 0.05, reading C from the curves, gives $C=1.064$. From formula (42), $R_0=5.24$ ohms per mi. which makes the resistance $R=5.53$ ohms per mi.

Abstracts of Bell System Technical Papers Not Appearing in this Journal

Voice-Frequency Carrier Telegraph System for Cables.[1] B. P. Hamilton, H. Nyquist, M. B. Long and W. P. Phelps. Carrier telegraph systems using frequencies above the voice range have been in use for a number of years on open-wire lines. These systems, however, are not suitable for long toll cable operation because cable circuits greatly attenuate currents of high frequencies. The system described in this paper uses frequencies in the voice range and is specially adapted for operation on long four-wire cable circuits ten or more telegraph circuits being obtainable from one four-wire circuit. The same carrier frequencies are used in both directions and are spaced 170 cycles apart. The carrier currents are supplied at each terminal station by means of a single multi-frequency generator.

Metallic Polar-Duplex Telegraph System for Long Small-Gage Cables.[2] John H. Bell, R. B. Shanck, and D. E. Branson. In connection with carrying out the toll-cable program of the Bell System, a metallic-circuit polar-duplex telegraph system was developed. The metallic-return type of circuit lends itself readily to the cable conditions, its freedom from interference allowing the use of low potentials and currents so that the telegraph may be superposed on telephone circuits. The new system represents an unusual refinement in direct current telegraph circuits, the operating current being of the same order of magnitude as that of the telephone circuits on which the telegraph is superposed.

The following are some of the outstanding features of the present system. Sensitive relays with closely balanced windings are employed in the metallic circuit, and "vibrating circuits" are provided for minimizing distortion of signals. Repeaters are usually spaced about 100 miles apart. Thirty-four-volt line batteries are used and the line current is four or five milli-amperes on representative circuits. Superposition is accomplished by the compositing method which depends upon frequency discrimination, the telegraph occupying the frequency range below that of the telephone. New local-circuit arrangements have been designed, employing polar relays for repetition of the signals; these arrangements are suitable for use in making up circuits in combination with carrier-current and ground-return polar-duplex telegraph sections. New forms of mounting are em-

[1] Journal A. I. E. E., Vol. 44, p. 213, 1925.

[2] Presented at the mid-winter convention of the A. I. E. E., Feb., 1925.

ployed in which a repeater is either built as a compact unit or is made up of several units which are mounted on I-beams, and subsequently interconnected. In the latter case the usual arrangements for sending and receiving from the repeater are omitted, and a separate "monitoring" unit provided for connection to any one of a group of repeaters.

The metallic system is suitable for providing circuits up to 1,000 miles or more in length, the grade of service being better than that usually obtained from ground-return circuits on open-wire lines for such distances. About 55,000 miles of this type of telegraph circuit are in service at present.

Polarized Telegraph Relays.[3] J. R. Fry and L. A. Gardiner. This paper discusses two forms of polarized telegraph relay which have been developed by the Bell System for metallic telegraph circuits and for carrier current telegraph circuits. Both relays are of the same general construction except that one is more sensitive and carries an auxiliary accelerating winding. The more sensitive relay is required to operate on reversals of line current of one milliampere, and at the same time retain its adjustment over long periods and faithfully and accurately repeat signals. It is interesting to note that under average conditions the ratio of power controlled by the contact circuit to that required by the line windings is about 5,000 to one. The parts entering into the magnetic circuit of this relay except for a permanent magnet, are made of the new magnetic alloy (permalloy) recently developed in the Bell Telephone Laboratories. Permalloy lends itself to use in this relay because of its high permeability and very small residual effects. The design of the relay armature and the support for the moving contacts is such that contact chatter is practically eliminated. Photo-micrograms showing practically no destructive action are given of the contacts of a relay which was in continuous service for 8½ months, during which time each contact made and broke its circuit approximately 45,000,000 times.

Supervisory Systems for Remote Control.[4] J. C. Field. With the great growth in power distributor systems and especially with the advent of the automatic substations with no attendant there has arisen need for a supervisory system to indicate to the central load dispatcher the position or operating condition of each important power unit in the outlying stations and also to give him means to operate promptly these power units when desired.

[3] Journal A. I. E. E., Vol. 43, p. 223, 1925.

[4] Electrical Communications, Vol. 3, pp.127–133, 1924.

By the turning of a key the dispatcher can open or close any switch or circuit breaker, start or stop any of the machines and receive back almost instantly a visual and continuous signal of a red or green lamp. The present systems provide in effect a key and two lamps, one red, one green, for each unit supervised mounted in easy access of the dispatcher.

Two main systems known as the distributor supervisory and the selector supervisory have been developed to meet the varying conditions of service.

The distributor system is recommended when there is a large number of units to be supervised in a given station. It consists essentially of two motor-driven distributors, one in each station, running in synchronism. Brushes on each distributor pass over corresponding segments of two sets of 50 segments at the same instant. Thus by means of only four connecting wires between the stations the control and continuous indication of 50 power units is possible.

The selector system is recommended when there is only a few switches to be supervised in a single station or in several stations located some distance apart. It consists essentially of hand operated keys to send predetermined codes of impulses to operate selectively step by step selectors at the distant stations. After the selector has operated the power unit, an auxiliary contact on this unit operates a motor-driven key to send coded impulses to operate a selector at the dispatcher's station to indicate the condition of the unit by lighting a red or green lamp. Several stations can be supervised over the same three-line wires.

The dispatcher, by looking at the lamps on his control board, can thus tell at all times the electrical and mechanical conditions at all points in the system and has means to change the operating conditions at any substation according to the demand for power.

Note on Dr. Louis Cohen's Paper on Alternating Current Cable Telegraphy.[5] L. A. MacColl. This is a criticism of two papers which were published in the Journal of the Franklin Institute by Dr. Louis Cohen. It is shown that Cohen's development of the theory of cable telegraphy contains many defects and errors, and in particular that his criticisms of H. W. Malcolm's book, "The Theory of the Submarine Telegraph and Telephone Cable," are without foundation.

Telephone Circuit Unbalances, Determination of Magnitude and Location.[6] L. P. Ferris and R. G. McCurdy. This paper dis-

[5] Journal of the Franklin Institute, Vol. 199, p. 99, 1925.
[6] Journal A. I. E. E., Vol. 43, p. 1133, 1924.

cusses the effects of unbalances of telephone circuits on noise and crosstalk, and describes methods for detecting the presence of these unbalances and locating them when detected. The maintenance of telephone circuits in a high state of efficiency with respect to balance is important since unbalances contribute to crosstalk between telephone circuits and to noise when such circuits are involved in inductive exposures. Different types of unbalances are included and their effects under different conditions of energization of the unbalanced circuit and neighboring conductors are discussed. Methods are described for determining:

(1) The general condition of circuits with respect to balance by crosstalk measurements from their terminals.

(2) The approximate location of unbalances along a line by measurements over a range of frequencies with a bridge at one end of the line.

(3) The final location of unbalances by field measurements with an unbalance detector which may be operated by a lineman and which usually does not require interruption of telephone service, except momentarily.

Toll circuit office unbalances are briefly discussed and a special bridge for detecting and measuring the unbalances of composite sets is described. A mathematical treatment of the bridge method for locating unbalances and a discussion of the necessity of terminating the circuits involved in the tests in their characteristic line impedances are given in an appendix. The methods and apparatus described are widely used in the Bell System and afford operating telephone companies means for maintaining their circuits in the condition of minimum practicable unbalance.

The Theory of Probability and Some Applications to Engineering Problems.[7] E. C. MOLINA. The purpose of this paper is to suggest a wider recognition by engineers of a body of principles which, in its mathematical form, is a powerful instrument for the solution of practical problems. Certain fundamental principles of the theory of probabilities are stated and applied to three problems from the field of telephone engineering.

Note on the Least Mechanical Equivalent of Light.[8] HERBERT E. IVES. In this paper the value for the brightness of the black body at the melting point of platinum recently obtained by the writer is

[7] Journal A. I. E. E., Vol. 44, p. 122, 1925.

[8] Journal of the Optical Society of American and Rev. of Scientific Instruments, Vol. 10, No. 3, March, 1925, p. 289.

used to find a value for the least mechanical equivalent of light using the latest values for the black body constants and the melting point of platinum. The spectral luminous efficiency curve obtained by Tyndall and Gibson is employed. It is found that over the entire range of probable values of the black body constants, the values for the least mechanical equivalent of light may be plotted as a straight line in terms of $\frac{C_2}{T}$ so that the present computations may be expressed in a simple equation in which any desired values of the black body constants may be inserted. Using the latest values the least mechanical equivalent of light is found to be .00161 watts per lumen. This is practically identical with the value obtained by using the author's earlier experimental determination using the monochromatic green mercury light, when combined with the Gibson and Tyndall luminous efficiency curve.

Photoelectric Properties of Thin Films of Alkali Metals.[9] HERBERT E. IVES. The thin films of alkali metals which deposit spontaneously on clean metal surfaces in highly exhausted inclosures are studied. The alkali metals, sodium, potassium, rubidium, and caesium, in the thin film form all exhibit, to a striking degree, the selective photoelectric effect first discovered in sodium-potassium alloy. Experiments on varying the thickness of the deposited film show that the selective effect only occurs at a certain stage of the film's development; for very thin films the selective effect is absent, and it disappears again for thick layers of the pure alkali metal. The wavelength maxima of emission previously ascribed to the selective effect in the pure alkali metals on the basis of observations with rough or colloidal surfaces are absent in these thin films.

The Normal and Selective Photoelectric Effects in the Alkali Metals and Their Alloys.[10] HERBERT E. IVES and A. L. JOHNSRUD. The photoelectric currents from specular surfaces of molten sodium, potassium, rubidium, and caesium, and their alloys are studied at various angles of incidence for the two principal planes of polarization. The selective photoelectric effect is clearly exhibited only in the case of the liquid alloy of sodium and potassium. Wave-length distribution curves show maxima of emission, which are usually, but not always, most pronounced for light polarized with the electric vector parallel to the plane of incidence. The wave-length maxima previously assigned to the several elements are not confirmed; the

[9] Astrophysical Journal, Vol. LX, No. 4, November, 1924.
[10] Astrophysical Journal, Vol. LX, No. 4, November, 1924.

maxima vary in position for the same element with the condition and mode of preparation of the surface.

Theory of the Schroteffekt.[11] T. C. Fry. The current from a vacuum tube is composed of discrete particles of electricity which emerge according to no regular law but in an accidental, statistical fashion. The current therefore fluctuates with time. If the fluctuations are amplified sufficiently they may be heard in a telephone receiver as "noise"—a type of noise which is due to the mechanism of electron emission itself and not to outside interference. This noise is called the "Schroteffekt."

The effect is of certain importance from the telephone standpoint, for it appears that signals, the intensity of which is lower than that of the accidental current fluctuations, can never be rendered intelligible by vacuum tube amplification since the noise due to the statistical fluctuations of space current would be amplified to the same extent and would mask the signals. Fortunately, however, the effect is much less pronounced under operating conditions than it is under the conditions which are most favorable for laboratory study. This is due to the fact that the presence of space charge under operating conditions smooths out the electron stream to a very material extent, and thus reduces the tube noise. The limitation imposed upon amplification is therefore not serious.

The present paper deals with what we have termed "laboratory conditions" as distinct from "operating conditions." Its principal result, arrived at by theoretical consideration, is: That if the electrons are emitted independently of one another the intensity of the noise in the measuring instrument is

$$S = \nu \bar{w}_1,$$

where ν is the number of electrons emitted per unit time and $\bar{w}_1$ is the average over all electrons of the energy that each would have caused to be dissipated in the measuring device if not other had ever been emitted.

When this formula is applied to the type of simply tuned circuit that was considered by earlier writers, it leads to substantially the the same results as they had obtained. It is more general than these earlier results, however, and rests on less questionable methods of derivation. It is, in fact, more general than the problem of the Schroteffekt itself and applies equally well to the absorption of energy from any type of accidental disturbance which satisfies the condition that the individual electromotive impulses occur inde-

[11] Journal of Franklin Institute, Vol. 199, p. 203, 1925.

pendently of one another. Static in radio telephony and certain types of crosstalk probably satisfy these conditions.

The Transmission Unit.[12] R. V. L. HARTLEY. The Bell System has recently adopted a new transmission unit, abbreviated *TU*, for expressing those quantities which heretofore have been expressed in miles of standard cable, or in Europe in terms of the βl unit. It is shown that units of this type measure the logarithm of a ratio, and that the present art requires that this ratio be that of two amounts of power. Any of the proposed units may be so defined. Their essential difference is in the ratio chosen to correspond to one unit. The ratio chosen for the *TU*, $10^{0.1}$, makes it nearly the same in size as the 800-cycle mile, which has advantages. It also facilitates the use of common logarithms in preference to natural logarithms for which the ratio e of the βl unit is adapted. A distortionless reference system calibrated in *TU* is discussed, and conversion tables for the various units are given.

The Thermionic Work Function of Oxide Coated Platinum.[13] C. DAVISSON and L. H. GERMER. Measurements of the thermionic work function of pure platinum coated with oxides of barium and strontium have been made simultaneously by two methods for the same segment of a uniformly heated filament. The theory of the measurements and the experimental arrangements are the same as used in an earlier experiment on the thermionic work function of pure tungsten.[14] Filament temperatures accurate to $\pm 5°$, were found from the resistance of the filament at 0° C. in conjunction with the temperature coefficients of resistance. (1) In the Calorimetric method the equivalent voltage of the work function was computed from the sudden voltage change resulting from switching off the space current, due to the cooling effect of the emission. The determination was much more difficult that in the case of the tungsten filament, and measurements were made at the signle temperature, 1064° K. At this temperature the work function ϕ was found to be equal to $1.79 \pm .03$ volts. (2) In the temperature variation method it was found that, after the temperature had been changed suddenly from one value to another, the emission changed approximately exponentially from an initial value to a final steady value. The half value period of this change varied from a few seconds at high temperature to over a quarter of an hour at low temperature. Interpreting this

[12] Electrical Communications, July, 1924. London Electrician, January 16 and 23, 1925.

[13] Physical Review, Vol. 24, p. 666, 1924.

[14] Davisson and Germer, Phys. Rev., 20, 300 (1922).

phenomenon as due to a progressive and reversible change of the character of the filament with temperature, the initial emissions after temperature changes from 1064° K. were used to determine the b constant of Richardson's equation corresponding to the equilibrium character of the filament at 1064° K, and similar measurements were made for the b constant corresponding to the character of the filament at 911° K. The two determinations lead, through the relationship $\phi = bk/e$, to 1.79 volts and 1.60 volts for the corresponding values of ϕ. For 1064° K, then, the two methods give values for ϕ in agreement. The measurements are, however, not sufficiently accurate to give any indication whether or not an electron within the metal possesses the thermal energy $3kT/2$. The various corrections made and possible errors are thoroughly discussed. It is pointed out that if the transition from the equilibrium state at one temperature to that at another had occurred so rapidly as to avoid observation, a disagreement of 25 per cent. between the values of ϕ given by the two methods would have been obtained which might have been misinterpreted.

Contributors to this Issue

Herbert E. Ives, B.S., University of Pennsylvania, 1905; Ph.D., Johns Hopkins, 1908; assistant and assistant physicist, Bureau of Standards, 1908–09; physicist, Nela Research Laboratory, Cleveland, 1909–12; physicist, United Gas Improvement Company, Philadelphia, 1912–18; U. S. Army Air Service, 1918–19; research engineer, Western Electric Company (Bell Telephone Laboratories), 1919 to date. Dr. Ives' work has had to do principally with the production, measurement and utilization of light.

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.

Ralzemond D. Parker, B.S., University of Michigan, 1905; M.S., 1906; instructor in Electrical Engineering, University of Michigan, 1906–09; assistant professor, 1909–13; Engineering Department, American Telephone and Telegraph Company, 1913–19; Department of Development and Research, 1919– . Mr. Parker's work has related particularly to telegraphy, included the development of printing telegraph apparatus, carrier, and metallic circuit systems for fine wire cables.

A. B. Clark, B.E.E., University of Michigan, 1911; American Telephone and Telegraph Company, Engineering Department, 1911–19; Department of Development and Research, 1919—. Mr. Clark's work has been connected with toll telephone and telegraph systems.

H. W. Nichols, B.S., 1908, E.E., 1911, Armour Institute of Technology; M.S., 1909, Ph.D., 1918, University of Chicago; Assistant Professor of Electrical Engineering, Armour Institute of Technology, 1909–11; Engineering Department, Western Electric Company (Bell Telephone Laboratories), 1911—. Since 1916 Mr. Nichols has been in charge of the laboratories research in radio communication.

J. C. Schelleng, A.B., 1915; instructor in physics, Cornell University, 1915–18; Engineering Department, Western Electric Company (Bell Telephone Laboratories), 1919—. Since 1918, Mr. Schelleng has been engaged in research in radio communication.

T. C. Smith, B.S., Purdue University, 1910; Plant Engineering, New York Telephone Company, 1910–14; engineering construction of high tension lines and municipal electric light plants, 1915; Outside Plant Engineering, New York Telephone Company, 1916–19; Automotive Engineering, New York Telephone Company, 1919–21; Automotive and Construction Apparatus Engineering, American Telephone and Telegraph Company, 1921–.

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.

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, 1917–24; Bell Telephone Laboratories, Inc., 1925—. Mr. Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

Sallie Pero Mead, A.B., Barnard College, 1913; M.A., Columbia University, 1914; American Telephone and Telegraph Company, Engineering Department, 1915–19; Department of Development and Research, 1919–. Mrs. Mead's work has been of a mathematical character relating to telephone transmission.

Courtesy of "Electrician," London

THE LATE OLIVER HEAVISIDE, F.R.S.

The Bell System Technical Journal

July, 1925

Oliver Heaviside

By F. GILL

ALTHOUGH abler pens[1] have expressed appreciation of the late Oliver Heaviside, it is perhaps permissible for an English telephone engineer to present a note regarding him. Of his life-history not very much is known; but he may have been influenced in his choice of a career by the fact that he was a nephew of the famous telegraph engineer Sir Charles Wheatstone. Heaviside was born in London on May 13, 1850; he entered the service of the Great Northern Telegraph Company, operating submarine cables, and he remained in that service, at Newcastle-on-Tyne, until 1874. While he was with the Telegraph Company, he published in 1873 a paper showing the possibility of quadruplex telegraphy.

At the age of about 24, owing, it is suggested, to increasing deafness, he left the service of that Company and took up mathematical research work. How he acquired his mathematical training does not seem to be known;[2] perhaps he was self-taught,—in some of his Papers he implies it. By whatever means he mastered the principles, it is evident that he was an ardent student of Maxwell, for constantly in Heaviside's own writing runs a vein of appreciation of Maxwell. For some time he lived in London, then he moved to Paignton in Devonshire; his Electrical Papers are written from there, and he died at the neighboring town of Torquay on February 4, 1925, in his 75th year.

That is about all the personal history at present available, and yet it gives a clue to a dominant note in his character, viz., reluctance to come into prominence, originating, perhaps, in a kind of shyness, which ultimately led to the recluse state. It is strange that so remarkable an investigator should, in his earlier manhood, have convinced so few, notwithstanding the fact that his voluminous writings made his name well known. It must, however, be remembered that his articles were very difficult, even for advanced mathematicians to follow, for he used a system of mathematics which, at that time

[1] *The Electrician*, Vol. XCIV, p. 174, by Sir Oliver Lodge, F.R.S., O.M. *Nature*, Vol. 115, p. 237, by Dr. Alex. Russell, F.R.S.

[2] Was he the youth with the frown in the library? He says he "then died," but also says "he was eaten up by lions." (E.M.T., Vol. III, pp. 1 and 135.)

was unusual. Whatever the cause, the fact remains that until about the year 1900 few engineers understood him.

Coming to his work, what was it that Heaviside did, and upon what does his fame rest? That is too large a subject for a telephone engineer to answer fully, but as regards communication engineering something may be said. His great achievement was the discovery of the laws governing the propagation of energy in circuits. He recognized the relationship between frequency and distortion; he illustrated it by numerical examples, and he showed what was required to make a "distortionless circuit." Further, he showed the effects of "attenuation" and the result of "inductance" (these words were his own coinage) in improving telephony. He also explained how the inductance of circuits could be increased; he suggested the use of continuous loading, of lumped inductance in the form of coils, and he pointed out the difficulty of obtaining sufficiently low resistance in such coils. He investigated the effect of sea and land and the upper atmosphere on the propagation of radio energy and how it was that this energy could be transmitted over the mountain of earth intervening between two distant places.

His activity in these matters can best be illustrated by extracts from his writings, as follows:

In his "Electrical Papers," Vol. II, written in 1887, p. 164, he gives numerical examples of frequency distortion and of its correction, and says:

> "It is the very essence of good long distance telephony that inductance should *not* be negligible."

In his "Electromagnetic Theory," Vol. I, published in 1893, he considers in Section 218, p. 441.

> "various ways, good and bad, of increasing the inductance of circuits"

He suggests, page 445, the use of

> ". . . inductance in isolated lumps. This means the insertion of inductance coils at intervals in the main circuit. That is to say just as the effect of uniform leakage may be imitated by leakage concentrated at distinct points, so we should try to imitate the inertial effects of uniform inductance by concentrating the inductance at distinct points. The more points the better, of course . . . The Electrical difficulty here is that inductance coils have resistance as well, and if this is too great the remedy is worse than the disease.

. To get large inductance with small resistance, or, more generally, to make coils having large time constants, requires the use of plenty of copper to get the conductance, and plenty of iron to get the inductance, employing a properly closed magnetic circuit properly divided to prevent extra resistance and cancellation of the increased inductance . . . This plan . . . is a straightforward way of increasing the *L* largely without too much increasing the resistance and may be worth working out and development. But I should add that there is, so far, no direct evidence of the beneficial action of inductance brought about in this way."

In "Electrical Papers," Vol. II, p. 311, he deals with reflected waves, and on page 347 he says:

" . . . but the transmitter and the receiving telephone distort the proper signals themselves. The distortion due to the electrical part of the receiver may, however, be minimized by a suitable choice of its impedance.

"Electromagnetic Theory," Vol. I, p. 404:—

"We have seen that there are four distinct quantities which fundamentally control the propagation of 'signals' or disturbances along a circuit, symbolized by *R*, *K*, *L*, and *S*, the resistance, external conductance, inductance, and permittance;"

"Electromagnetic Theory," Vol. I, p. 411:—

"It is not merely enough that signals should arrive without being distorted too much; but they must also be big enough to be useful . . . Nor can we fix any limiting distance by consideration of distortion alone. And even if we could magnify very weak currents, say a thousandfold, at the receiving end, we should simultaneously magnify the foreign interferences. In a normal state of things interferences should be only a small fraction of the principal or working current. But if the latter be too much attenuated, the interferences become relatively important, and a source of very serious distortion. We are, therefore, led to examine the influence of the different circuit constants on the attenuation, as compared with their influence on the distortion."

"Electrical Papers," Vol. II, p. 402:—

"I was led to it (the distortionless circuits), by an examination of the effect of telephones bridged across a common circuit (the proper place for intermediate apparatus, removing their impedance) on waves transmitted along the circuit."

With regard to Radio Communication, one extract must suffice writing on The Electric Telegraph in June, 1902, for the Encyclopedia Britannica, he says,—"Electromagnetic Theory," Vol. III, p. 335:—

"There is something similar in 'wireless' telegraphy. Sea water, though transparent to light, has quite enough conductivity to make it behave as a conductor for Hertzian waves, and the same is true in a more imperfect manner of the earth. Hence the waves accommodate themselves to the surface of the sea in the same way as waves follow wires. The irregularities make confusion, no doubt, but the main waves are pulled round by the curvature of the earth, and do not jump off. There is another consideration. There may possibly be a sufficiently conducting layer in the upper air. If so, the waves will, so to speak, catch on to it more or less. Then the guidance will be by the sea on one side and the upper layer on the other. But obstructions, on land especially, may not be conducting enough to make waves go round them fairly. The waves will go partly through them."

Probably due to his long seclusion, his approach to certain subjects was rather critical. At one time I tried to get a portrait of him for the Institution of Electrical Engineers, but failed;—he did not wish to have his photograph exhibited, he thought that "one of the worst results (of such exhibition) was that it makes the public characters think they really are very important people, and that it is therefore a principle of their lives to stand upon doorsteps to be photographed."

On another occasion when I sent him a copy of an article by a distinguished telephone engineer on "The Heaviside Operational Calculus," he replied that he had "looked through the paper . . . with much interest, to see what progress is being made with the academical lot, whom I have usually found to be very stubborn and sometimes wilfully blind."

Some have held that Heaviside was not recognized as he ought to have been. This was probably the case some time ago, but not in recent years. The same is true of many very great men who were much in advance of their time, for the English have the national characteristic that they do not make much fuss about their great men. So if Heaviside suffered, he shared this experience in common with other pioneers who deserved higher recognition. See, for example, what Heaviside himself said about one of these, in a footnote in "Electromagnetic Theory," Vol. III, p. 89:

"George Francis Fitzgerald is dead. The premature loss of a man of such striking original genius and such wide sympathies will be considered by those who knew him and his work to be a national misfortune. Of course, the 'nation' knows nothing about it, or why it should be so."

During the last 20 years or more, the significance and luminous quality of the work of Heaviside has been increasing by acknowledged mathematicians and by practical telephone, telegraph and radio engineers. To other electrical engineers his treatment of wave-transmission has not yet appealed quite so strongly.

Probably his first recognition came from his contribution to the problem—"Electromagnetic Induction and its Propagation" in the *Electrician*. It appeared as a series of articles between January, 1885 and December, 1887. His "Electrical Papers" were written at various times and were published in two volumes in 1892. Then followed his three volumes on "Electromagnetic Theory"—on the basis of the *Electrician* articles published in 1893, 1899 and 1912. He also wrote, in 1902, the article on the "Theory of the Electric Telegraph" in the "Encyclopedia Britannica."

In 1891, the Royal Society made him a Fellow. In 1899, the American Academy of Arts and Sciences elected him an Honorary Member. In 1908 the Institution of Electrical Engineers did the same, followed by the American Institute of Electrical Engineers in 1917. The Literary and Philosophical Society of Manchester also elected him an Honorary Member. He was an Hon. Ph.D. of the University of Gottingen, and in 1921, the Institution of Electrical Engineers conferred upon him the highest award in their gift—the Faraday Medal. He was the first recipient of this Medal which was established to commemorate the 50th anniversary of the founding of the original Society of Telegraph Engineers and of Electricians, and since then the medal has been bestowed upon Sir Charles Parsons, Dr. S. Z. de Ferranti, and Sir J. J. Thomson.

From time to time there were reports of his living in great poverty, and attempts were made to help him. These reports lacked proportion, but it is true he had not much money and perhaps still less comfort; he was a difficult man to help. Towards the end of his life he received from the British Government a Civil Pension. His independent character rendered it necessary that offers of assistance should be tactfully made and apparently this was not always the case, as I believe help was sometimes refused; but there were those who succeeded. Another difficulty was his unconventional mode of living which caused him, in his last years, to live as a recluse, cooking and looking after his house alone.

Just what other work Heaviside did, in addition to his published writings, is not at present known to me. I believe he left a good deal of manuscript, but whether it is in such a state that it could be completed by another, I do not know. Let me conclude this note by an extract from his last chapter of his last book, "Electromagnetic Theory," Vol. III, page 519:—

> "As the universe is boundless one way, towards the great, so it is equally boundless the other way, towards the small; and important events may arise from what is going on in the inside of atoms, and again, in the inside of electrons. There is no energetic difficulty. Large amounts of energy may be very condensed by reason of great forces at small distances. How electrons are made has not yet been discovered. From the atom to the electron is a great step, but is not finality.
>
> "Living matter is sometimes, perhaps generally, left out of consideration when asserting the well-known proposition that the course of events in the physical world is determined by its present state, and by the laws followed. But I do not see how living matter can be fairly left out. For we do not know where life begins, if it has a beginning. There may be and probably is no ultimate distinction between the living and the dead."

The Loaded Submarine Telegraph Cable[1]

By OLIVER E. BUCKLEY

SYNOPSIS: With an increase of traffic carrying capacity of 300%, over that of corresponding cables of the previous art, the New York-Azores permalloy-loaded cable marks a revolution in submarine cable practice. This cable represents the first practical application of inductive loading to transoceanic cables. The copper conductor of the cable is surrounded by a thin layer of the new magnetic material, permalloy, which serves to increase its inductance and consequently its ability to transmit a rapid succession of telegraph signals.

This paper explains the part played by loading in the operation of a cable of the new type and discusses some of the problems which were involved in the development leading up to the first commercial installation. Particular attention is given to those features of the transmission problem wherein a practical cable differs from the ideal cable of previous theoretical discussions.

Brief mention is made of means of operating loaded cables and the possible trend of future development.

Permalloy Loading

THE announcement on September 24, 1924, that an operating speed of over 1,500 letters per minute had been obtained with the new 2,300 mile New York-Azores permalloy-loaded cable of the Western Union Telegraph Company, brought to the attention of the public a development which promises to revolutionize the art of submarine cable telegraphy. This announcement was based on the result of the first test of the operation of the new cable. A few weeks later, with an improved adjustment of the terminal apparatus, a speed of over 1,900 letters per minute was obtained. Since this speed represents about four times the traffic capacity of an ordinary cable of the same size and length, it is clear that the permalloy-loaded cable marks a new era in transoceanic communication.

The New York-Azores cable represents the first practical attempt to secure increased speed of a long submarine telegraph cable by inductive loading and it is the large distributed inductance of this cable which is principally responsible for its remarkable performance. This inductance is secured by surrounding the conductor of the cable with a thin layer of permalloy. Fig. 1 shows the construction of the deep sea section of the cable. In appearance it differs from the ordinary type of cable principally in having a permalloy tape 0.005 inch thick and 0.125 inch wide, wrapped in a close helix around the stranded copper conductor.

Permalloy, which has been described by Arnold and Elmen,[2] is an alloy consisting principally of nickel and iron, characterized by very

[1] Presented before the A. I. E. E., June 26, 1925.

[2] *Jour. Franklin Inst.*, Vol. 195, pp. 621–632, May 1923; *B. S. T. J.*, Vol. II, No. 3, p. 101.

high permeability at low magnetizing forces. The relative proportion of nickel and iron in permalloy may be varied through a wide range of additional elements as, for example, chromium may be added to secure high resistivity or other desirable properties. On account

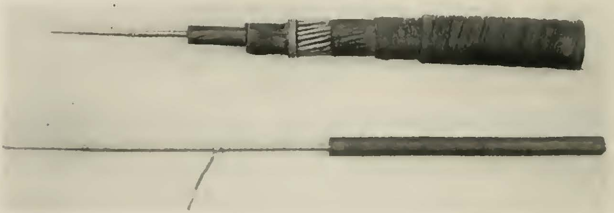


Fig. 1—Permalloy-Loaded Cable. Above, section of deep sea type showing construction. Below, section of core showing permalloy tape partly unwound.

of its extremely high initial permeability a thin layer of permalloy wrapped around the copper conductor of a cable greatly increases its inductance even for the smallest currents.

In the case of the New York-Azores cable the permalloy tape is composed of approximately 78½% nickel and 21½% iron and gives the cable an inductance of about 54 millihenries per nautical mile. An approximate value of the initial permeability of the permalloy in that cable may be got by assuming the helical tape replaced by a continuous cylinder of magnetic material of the same thickness.[3] This material would have to have a permeability of about 2,300 to give the observed inductance. A better appreciation of the extraordinary properties of the new loading material may be obtained by comparing this permeability with that which has previously been obtained with iron as the loading material. The Key West-Havana telephone cables are loaded with 0.008 inch diameter soft iron wire. The permeability of this wire, which was the best which could be obtained commercially when that cable was made, is only about 115,

[3] The true initial permeability is slightly higher. To compute it, account must be taken of the fact that, contrary to what has been sometimes assumed, the magnetic lines of induction in the tape do not form closed loops around the wire but tend to follow the tape in a helical path. The pitch of the helical path of the lines of induction is slightly less than that of the permalloy tape with the result that a line of induction takes a number of turns around the conductor, then crosses an airgap between two adjacent turns of tape and continues along the tape to a point where it again slips back across an airgap. O. E. Buckley, British Patent No. 206,104, March 27, 1924, also K. W. Wagner, E.N.T., Vol. 1, No. 5, p. 157, 1924.

or approximately one-twentieth that of the permalloy tape of the New York-Azores cable.

Problems Encountered

The proposal to use permalloy loading to increase the speed of long telegraph cables was one outcome of an investigation undertaken by the author soon after the war to determine whether some of the new methods and materials developed primarily for telephony might not find important application to submarine telegraphy. In the subsequent development of the permalloy loaded cable a large number of new problems, both theoretical and practical, had to be solved before the manufacture of a cable for a commercial project could be undertaken with reasonable assurance of success. The problems encountered were of three principal kinds. First was that of the transmission of signals over a cable having the characteristics of the trial conductors made in the laboratory. Although the theory of transmission over a loaded cable had been previously treated by others, the problem considered had been that of an ideal loaded cable with simple assumptions as to its electrical constants and without regard to the practical limitations of a real cable. The second class of problems had to do with the practical aspects of design, manufacture and installation. In this connection an extensive series of experiments was conducted to determine the means required to secure at the ocean bottom the characteristics of the laboratory samples on which the transmission studies were based. Among the numerous problems which arose in this connection were those concerned with protecting the copper conductor from any possible damage in the heat-treating operation which was necessary to secure the desired magnetic characteristics, and those concerned with protecting the strain-sensitive permalloy tape from being damaged by submerging the cable to a great depth. The third class of problem had to do with terminal apparatus and methods of operation. The prospective speed of the new cable was quite beyond the capabilities of standard cable equipment and accordingly new apparatus and operating methods suited to the loaded cable had to be worked out. In particular it was necessary to develop and construct instruments which could be used to demonstrate that the speed which had been predicted could actually be secured. The success of the investigations along all three lines is attested by the results which were obtained with the New York-Azores cable. Fig. 2 shows a section of cable recorder slip, the easily legible message of which was sent from

Horta, Fayal, and received at New York at a speed of 1,920 letters per minute.

It is principally with regard to the first of these classes of problems, that of the transmission of signals, that the following discussion is concerned. No attempt will be made here to discuss the details of

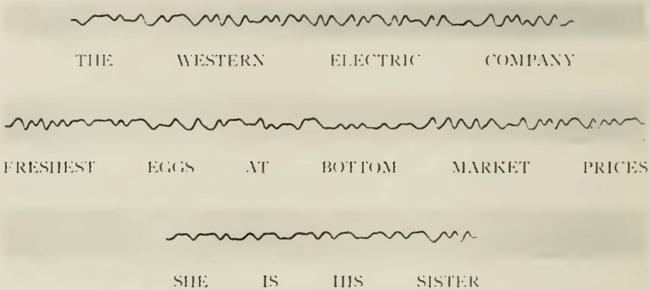


Fig. 2—Test Message. Western Union New York-Azores Permalloy-Loaded Cable. Sent from Horta (Azores) and received at New York, November 14, 1924. Speed—1920 letters per minute. Recorded with special high speed siphon recorder

design and development of the physical structure of the cable, nor will there be given a detailed description of the operating results or how they were obtained. These subjects must be reserved for later publication. It is desired in what follows to explain how inductive loading improves the operation of a submarine cable and to point out some of the problems concerned with the transmission of signals which had to be considered in engineering the first long loaded cable.

Factors Limiting Speed of Non-Loaded Cable

In order to understand the part played by loading in the transmission of signals it is desirable first to review briefly the status of the cable art prior to the introduction of loading and to consider the factors then limiting cable speed and the possible means of overcoming them. A cable of the ordinary type, without loading, is essentially, so far as its electrical properties are concerned, a resistance with a capacity to earth distributed along its length. Although it does have some inductance, this is too small to affect transmission at ordinary speeds of operation except on cables with extremely heavy conductors. The operating speed of a non-loaded cable is approximately inversely proportional to the product of the total resistance by the total capacity: that is,

$$S = \frac{k}{CRl^2},$$

where C is capacity and R resistance per unit length, and l is the length of the cable. The coefficient k is generally referred to as the speed constant. It is, of course, not a constant since it depends on such factors as terminal interference and method of operation, but is a convenient basis for comparing the efficiency of operation of cables of different electrical dimensions. As the technique of operating cables has improved the accepted value of k has increased, its value at any time being dependent on the factor then limiting the maximum speed obtainable. This factor has at times been the sensitiveness of the receiving apparatus, at other times the distortion of signals, and in recent years interference. During a great part of the history of submarine cable telegraphy distortion was considered the factor which limited the speed of operation of long cables and on this account most of the previous discussions of submarine cable transmission have been concerned principally with distortion and means for correcting it. As terminal apparatus was gradually improved means of correcting distortion were developed which practically eliminated distortion as an important factor in the operation of long cables. With distortion thus eliminated the speed was found to be limited principally by the sensitiveness of the receiving apparatus. This limit was, however, eliminated in turn by the development of signal magnifiers. During recent years, in which numerous cable signal magnifiers have been available and methods of correcting distortion have been understood, the only factor limiting cable speed has been the mutilation of the feeble received signals by interference. Most cables are operated duplex, and in these the speed is usually limited by interference between the outgoing and incoming signals. In cables operated simplex, and also in cables operated duplex where terminal conditions are unfavorable, speed is limited by extraneous interference which may be from natural or man-made sources and which varies greatly in different locations. The strength of the received current must in either case be great enough to make the signals legible through the superposed interference current. Owing to the rapidity with which the received signal amplitude is decreased as the speed of sending is increased, the limiting speed is quite sharply defined by the interference to which the cable is subject.

Means of Increasing Speed

With the speed of operation thus limited there were two ways in which the limiting speed could be increased: the interference could be reduced, or the strength of signals made greater. No great reduction in interference due to lack of perfect duplex balance could be expected, as balancing networks had already been greatly refined. Extraneous interference in certain cases could be reduced by the use of long, properly terminated sea-earths. The signal strength could be increased either by increasing the sending voltage or by decreasing the attenuation of the cable. However, with duplex operation nothing at all is gained by increasing the voltage in cases where lack of perfect duplex balance limits the speed, and with simplex operation any gain from raising the voltage is obtained at the cost of increased risk to the cable, the sending voltage being usually limited to about 50 volts by considerations of safety. The attenuation of the cable could be reduced and the strength of the signal increased by use of a larger copper conductor or by using thicker or better insulating material. None of these possible improvements, however, seemed to offer prospect of very radical advance in the art.

In telephony, both on land and submarine lines, an advantage had been obtained by adding inductance [3] in either of two ways, by coils inserted in series with the line or by wrapping the conductor with a layer of iron. The insertion of coils in a long deep-sea cable was practically prohibited by difficulties of installation and maintenance. Accordingly, only the second method of adding inductance, commonly known as Krarup or continuous loading, could be considered

[3] The idea of improving the transmission of signals over a line by adding distributed inductance to it originated with Oliver Heaviside in 1887 (Electrician, Vol XIX, p. 79, and Electromagnetic Theory, Vol. I, p. 441, 1893), who was the first to call attention to the part played by inductance in the transmission of current impulses over the cable. He suggested as a means for obtaining increased inductance the use of iron as a part of the conductor or of iron dust embedded in the gutta percha insulation. He also proposed inserting inductance coils at intervals in a long line. Other types of coil loading were proposed by S. P. Thompson (British Patent 22,304—1891, and U. S. Patents 571,706 and 571,707—1896), and by C. J. Reed (U. S. Patents 510,612 and 510,613—1893). M. I. Pupin (A. I. E. E. Trans., Vol. XVI, p. 93, 1899, and Vol. XVII, p. 445, 1900) was the first to formulate the criterion on the basis of which coil loaded telephone cables could be designed. Continuous loading by means of a longitudinally discontinuous layer of iron covering the conductor was proposed by J. S. Stone in 1897 (U. S. Patent 578,275). Breisig (E. T. Z., Nov. 30, 1899) suggested the use of an open helix of iron wire wound around the conductor and Krarup (E. T. Z., April 17, 1902) proposed using a closed spiral so that the adjacent turns were in contact. J. H. Cuntz (U. S. Patent 977,713 filed March 29, 1901) proposed another form of continuous loading. Recent general discussions of loaded telegraph cable problems have been given by Malcolm (Theory of Submarine Telegraph and Telephone Cable, London, 1917) and by K. W. Wagner (Elektr. Nachtr. Tech., Oct., 1924).

for a transoceanic telegraph cable and it is primarily with regard to continuous loading that the following discussion is concerned.

Effects of Loading

Most of the proposals to load telegraph cables have had the object of reducing or eliminating distortion, and accordingly most of the mathematical treatments of loading have been from that point of view. The reduction of distortion is, however, not the only benefit to be obtained from loading and, in fact, may not always be secured in the high speed operation of a loaded cable. The principal benefit of loading from the practical standpoint is to decrease the attenuation of the signals so that for a given frequency more current will be received or so that the minimum permissible current may be received with a greater speed of signalling. From the mathematical standpoint there are two ways of treating the problem of the loaded cable, first with regard to the transmission of a transient impulse, and second with regard to setting up steady alternating currents of definite frequency. In the ultimate analysis the solution of either problem can be got from the other. However, for practical purposes they are two distinct means of attack. Which should be used depends on the object to be secured. If one is concerned primarily with the effect of the cable on the wave shape of the signal transmitted over it, it is fairly obvious that the transient treatment has advantages. If, however, one is concerned only with the strength of the received signal, as is the case if there is assurance that the signal shape can in any event be corrected by terminal networks, then the steady state treatment is sufficient and much more convenient to apply. In the case of the real loaded cable the complete transient solution is extremely complex and the steady state treatment relatively simple. The solution of the transient problem of an ideal loaded cable is, however, very valuable to give a physical picture of how inductive loading aids the high speed transmission of signals.

The transient solution of the problem of an ideal heavily loaded cable has been worked out by Malcolm [4] and more rigorously by Carson [5], who have determined the curve showing the change of current with time at one end of the cable if a steady e.m.f. is applied at zero time between the cable and earth at the distant end. Such a curve is called an "arrival curve" and for an ideal loaded cable comprising only constant distributed resistance, capacity and inductance may have a form like that shown in Curve b of Fig. 3, which is to be

[4] Theory of the Submarine Telegraph and Telephone Cable, London, 1917.

[5] Trans. A. I. E. E., Vol. 38, p. 345, 1919.

compared with Curve a which is the arrival curve of a non-loaded cable. The straight vertical part of Curve b represents the "head" of the signal wave which has travelled over the cable at a definite speed and with diminishing amplitude. The definite head of the arrival curve is the most striking characteristic difference between the ideal loaded and the non-loaded cable. In the latter, as is evident from Fig. 3, the current at the receiving end starts to rise slowly almost as soon as the key is closed at the transmitting end. When an e.m.f. is applied to the sending end of the non-loaded cable a charge spreads out rapidly over the whole length, the receiving end charging up much more slowly than the sending end on account of the resistance of the intervening conductor. Hence, if a signal train consisting of rapidly alternating positive and negative impulses is applied to the sending end, the effect at the receiving end of charging the cable positively is wiped out by the succeeding negative charge before there has been time to build up a considerable positive potential and the successive alternating impulses thus tend to annul each other. In the loaded cable the effect of inductance is to oppose the setting up of a current and to maintain it once it has been established, and thus to maintain a definite wave front as the signal impulse travels over the cable. Hence, with inductive loading the strength and individuality of the signal impulses are retained and a much higher speed of signalling is possible. It should be noted that by speed of signalling is meant the rapidity with which successive impulses are sent and not the rate at which they travel over the cable. This speed of travel is actually decreased by the addition of inductance, about one-third of a second being required for an impulse to traverse the New York-Azores cable from end to end.

It should be noted that Curve b of Fig. 3 is for an ideal loaded cable in which the factors of resistance, capacity and inductance are constant. In a real loaded cable none of these factors are constant and the arrival curve cannot be simply and accurately computed. Even the capacity which is usually assumed as constant for real cables varies appreciably with frequencies in the telegraph range, and owing to the fact that gutta percha is not a perfect dielectric material its conductance, which is also variable with frequency, must be taken into account. Although the inductance of the cable is substantially constant for small currents of low frequency, it is greater for the high currents at the sending end of the cable on account of the increase of magnetic permeability of the loading material with field strength and is less at high frequencies than at low on account of the shielding effect due to eddy currents. The resistance is highly variable since it comprises, in addition to the resistance of the copper conductor, effective resistance due to eddy currents and hysteresis in the loading material, both of which vary with frequency and current amplitude. Furthermore, there is variable inductance and resistance in the return circuit outside the insulated conductor which must be

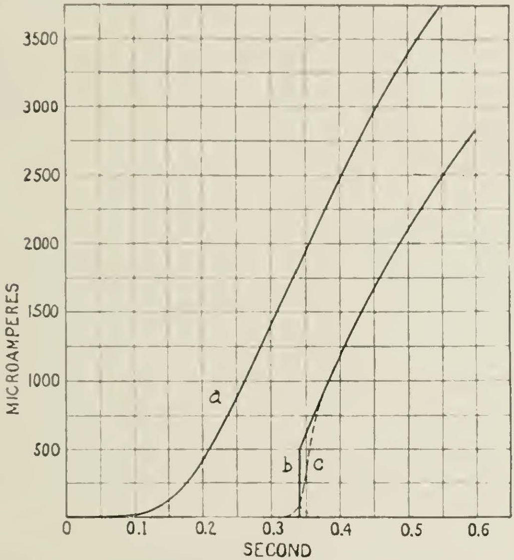


Fig. 3—Arrival Curves. a. Non-loaded cable. b. Ideal loaded cable. c. Real loaded cable (approximate)

taken into account. Although it is very difficult to compute the exact arrival curve of a cable subject to all of these variable factors, an approximate calculation in a specific case like that of the New York-Azores cable shows that the arrival curve has the general shape of Curve c of Fig. 3. It will be noticed that although this arrival curve lacks the sharp definite head, characteristic of the ideal loaded cable, it still has a relatively sharp rise and that the time required for the impulse to traverse the cable is not greatly different from that of the ideal loaded cable.

Although it is difficult to take exact account of the variable characteristics of the loaded cable in the solution of the transient problem, it is easy to take account of them in the steady state or periodic analysis by means of well-known methods. If a steady sinusoidal voltage, V_s, is applied at one end of the cable the resulting voltage, V_r, at the distant end will be given by the equation

$$V_r = kV_s\epsilon^{-Pl},$$

where l is the length, P, the propagation constant of the cable and k, a constant which depends on the terminal impedance and which is unity in case the cable is terminated at the receiving end in its so-called characteristic impedance. The propagation constant is given by the formula,

$$P = \sqrt{(R+ipL)(G+ipC)} = \alpha + i\beta,$$

where R is the resistance, L, the inductance, G, the leakance and C, the capacity per unit length and p is 2π times the frequency. The real part of the propagation constant, α, is called the attenuation constant and the imaginary part, β, the wave length constant. By separating α and β the amplitude and phase displacement of the received voltage relative to the sent voltage may be computed for any particular frequency and the behavior of a complex signal train may be worked out by analyzing it into its Fourier components and treating them separately. The phase shift is, however, of importance mainly as regards the shape of the received signals and their amplitude may, in general, be obtained from the attenuation constant alone. Thus if it is known that the signal shape can in any case be corrected by terminal networks there is no need to be concerned with more than the attenuation constant to compute the speed of the cable.

In the case of a cable of the permalloy loaded type, α is given with an approximation [6] sufficiently close for the purposes of this discussion by the equation,

$$\alpha = \frac{1}{2}\sqrt{\frac{C}{L}}\left(R + \frac{G}{C}L\right).$$

For the purpose of computing R it is convenient to separate it into its components, giving

$$\alpha = \frac{1}{2}\sqrt{\frac{C}{L}}\left(R_c + R_e + R_s + R_h + \frac{G}{C}L\right),$$

[6] For accurate computation of attenuation the complete formula for α must be used.

where R_c = copper resistance per unit length

R_e = eddy current resistance per unit length

R_s = sea return resistance per unit length

R_h = hysteresis resistance per unit length

The copper resistance R_c is that determined by a direct current measurement of the loaded conductor since the resistance of the loading tape is so high and its length is so great that the current flowing longitudinally through it may be safely neglected.

The eddy current resistance R_e is given approximately by the formula,

$$R_e = \frac{m\mu^2 l^3 f^2}{\rho\,(d-l)},$$

where l is the thickness or diameter of the loading tape or wire, d, the outside diameter of the loaded conductor, f, the frequency, ρ, the resistivity of the loading material, μ, its magnetic permeability and m, a constant which depends on the form of the loading material and is in general greater for tape than for wire loading. Although it is possible to compute a value of m, the value found in practice is always larger than the theoretical value which is necessarily based on simple assumptions and does not take into account such a factor as variation of permeability through the cross-section or length of the loading material. Accordingly it is necessary to determine m experimentally for any particular type of loaded conductor.

The sea-return resistance may be safely neglected in the computation of slow speed non-loaded cables, but it is a factor of great consequence in the behavior of a loaded cable. By sea-return resistance is meant the resistance of the return circuit including the effect of the armor wire and sea water surrounding the core of the cable. Although the exact calculation [7] of this resistance factor is too complex to be discussed here, the need for taking it into account may be quite simply explained. Since the cable has a ground return, current must flow outside the core in the same amount as in the conductor. The distribution of the return current is, however, dependent on the structure of the cable as well as on the frequencies involved in signalling. If a direct current is sent through a long cable with the earth as return conductor the return current spreads out through such a great volume of earth and sea water that the resistance of the return path is negligible. On the other hand if an alternating current is sent through the cable the return current tends to concentrate

[7] See Carson and Gilbert, *Jour. Franklin Inst.*, Vol. 192, p. 705, 1921; *Electrician*, Vol. 88, p. 499, 1922; *B. S. T. J.*, Vol. 1, No. 1, p. 88.

around it, the degree of concentration increasing with the frequency. With the return current thus concentrated the resistance of the sea water is of considerable consequence. It is further augmented by a resistance factor contributed by the cable sheath. This may be better understood by considering the cable as a transformer of which the conductor is the primary and the armor wire and sea water are each closed secondary circuits. Obviously the resistances of the secondary circuits of armor wire and sea water enter into the primary circuit and hence serve to increase the attenuation. The presence of the armor wires may thus be an actual detriment to the transmission of signals.

To take account of the hysteresis resistance, R_h, and also of the increased inductance and eddy current resistance at the sending end of the cable it is most convenient to compute the attenuation of the cable for currents so small that R_h may be safely neglected. The attenuation thus computed is that which would be obtained over the whole cable if a very small sending voltage were used. The additional attenuation at the sending end for the desired sending voltage may then be approximated by computing successively from the sending end the attenuation of short lengths of cable over which the current amplitude may be considered constant, the attenuations of separate lengths being added together to give the attenuation of that part of the cable in which hysteresis cannot be neglected. In this computation account must, of course, be taken of the increased inductance and eddy current resistance accompanying the higher currents at the sending end.

Having calculated or obtained by measurement the several resistance factors and knowing the capacity, leakance and inductance, the whole attenuation of a cable for any desired frequency may be computed and a curve drawn showing the variation of received current with frequency for a given sending voltage. This relation for a particular case is shown in Curve c of Fig. 4. Curve a shows for comparison the relation between frequency and received current of a non-loaded cable of the same size, that is, a cable having a conductor diameter the same as that of the loaded conductor and having the same weight of gutta percha. Curve b shows the behavior of an ideal loaded cable having the same inductance, capacity and d.c. resistance as the real loaded cable of Curve c, but in which the leakance and alternating current increments of resistance are assumed to be zero.

Now, if the level of interference through which the current must be received is known, the maximum speed of signalling for the loaded cable may be obtained from Curve c. It is that speed at which the highest frequency necessary to make the signals legible is received with sufficient amplitude to safely override the superposed interference. Just what the relation of that frequency is to the speed of signalling cannot be definitely stated, since it depends on the method of operation and code employed as well as on the desired perfection

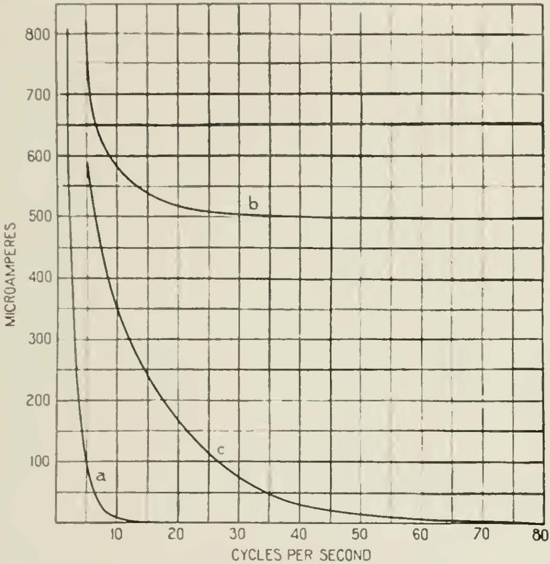


Fig. 4—Received Current vs. Frequency. a. Non-loaded cable. b. Ideal loaded cable. c. Real loaded cable

of signal shape. J. W. Milnor[5] has suggested that for cable code operation and siphon recorder reception a fair value is about 1.5 times the fundamental frequency of the signals, that is, the fundamental frequency when a series of alternate dots and dashes is being sent.

Remarks on the Design of Loaded Cables

By referring again to the equation for α, above, it can now be explained why high permeability is a necessary characteristic of the

[5] *Journal A. I. E. E.*, Vol. 41, p. 118, 1922. Transactions A. I. E. E., Vol. 41, p. 20, 1922.

loading material if a benefit is to be obtained from continuous loading. The addition of the loading material has two oppositely directed effects; on the one hand it tends to improve transmission by increasing the inductance and consequently decreasing the attenuation, and on the other hand it tends to increase the attenuation by increasing the effect of leakance and by the addition of resistance. Not only are the hysteresis and eddy-current factors of resistance added by the loading material but it must also be looked upon as increasing either the copper resistance or the capacity on account of the space it occupies. Generally it is more convenient to look upon the loading material as replacing some of the copper conductor in the non-loaded cable with which comparison is made, since by so doing all of the factors outside of the loaded conductor are unchanged. Now, if the loading material is to be of any benefit, the decrease in attenuation due to added inductance must more than offset the increase due to added resistance, including the added copper resistance due to the substitution of loading material for copper. In the limiting case the lowest permeability material which will show a theoretical advantage from this point of view is that which, as applied in a vanishingly thin layer, gives more gain than loss. For any particular size and length of cable there is a limiting value of permeability which will satisfy this condition, this limiting value being greater the longer the cable and the smaller the diameter of its conductor.[9] For transatlantic cables of sizes laid prior to 1923 the minimum initial permeability required to show an advantage is higher than that of any material known prior to the invention of permalloy. Actually a considerably higher permeability than this theoretical minimum was, of course, required to make loading an economic advantage since there are practical limits to the thickness of loading material and since the cost of applying it has also to be taken into account. Further, there are limits on methods of operation imposed by loading which necessitate still higher permeability to make loading worth while.

Since the addition of loading has two opposite tendencies in its effect on attenuation, the practical design of the cable must be based on a compromise between them. Thus, to secure the maximum gain from loading a cable of a given size, the loading material should be chosen of such a thickness that the gain due to increased inductance from a slight increase of thickness just offsets the loss due to increased resistance and dielectric leakance. In practice, of course, economic considerations of the cost of various thicknesses of loading must also be taken into account.

[9] See British Patent No. 184,774—1923, to O. E. Buckley.

In designing the New York-Azores cable some assumption had to be made as to the extraneous interference which would be encountered. Theoretical considerations led us to believe that the loaded cable would be no more subject to external interference than non-loaded cables. It even appeared that it would be less affected by some types of interference, for, owing to the shorter wave-length for a given frequency, a disturbance which affects a great many miles of cable simultaneously is less cumulative in its effect at the terminal of a loaded than a non-loaded cable. A reasonable assumption seemed to be that the total overall attenuation which could be tolerated for the loaded cable was at least as great as that which experience had shown to be permissible for simplex operation of non-loaded cables. This maximum permissible attenuation depends, of course, on conditions of terminal interference and no fixed value can be given as applicable to all cables. However, for average conditions of terminal interference in locations free from power line disturbances and where the cable lies in relatively deep water near to its terminal landing, a reasonable value of total attenuation constant for the fundamental frequency of cable code is about 10 (86.9 T.U.) for recorder operation and about 9 (78.2 T.U.) for relay operation. These were the approximate values assumed for the New York-Azores cable and later experience has demonstrated that they were well justified.

Distortion in Loaded Cables

Throughout all of the preceding discussion it has been assumed that the relation between attenuation and terminal interference would limit the speed of simplex operation rather than that distortion of signal shape would be the limiting factor. Although this is, in fact,[10] the case with non-loaded cables it was not self-evident as regards the loaded cable, and to make reasonably certain that the speed could be determined from the attenuation-frequency relation required a demonstration that the signal distortion of a real loaded cable could be corrected by suitable terminal apparatus. One of the merits long claimed for loading was that it would reduce distortion and, indeed, an ideal loaded cable with constant inductance and without magnetic hysteresis, eddy current loss, dielectric leakance and sea return resistance would have very little distortion and would give a speed limited only by terminal apparatus. However,

[10] Recent work of J. R. Carson (U. S. Patent 1,315,539—1919) and R. C. Mathes (U. S. Patent 1,311,283—1919) has shown that with the combined use of vacuum tube amplifiers and distortion correcting networks, distortion in non-loaded cables can be compensated to any desired degree.

a real loaded cable, the inductance of which varies with both current and frequency and in which all the above noted resistance factors are present, may give, and in general will give when operated at its maximum speed, greater distortion of signals than a non-loaded cable.

To solve the question of distortion on a purely theoretical basis required consideration of the transmission of a transient over the loaded cable. This was made extremely difficult by the existence of numerous possible causes of signal distortion, the effects of which could only be approximated in the solution of the transient problem. In addition to the distortion resulting from the rapid increase of attenuation with frequency due to the various sources of alternating current losses, distortion peculiar to the magnetic characteristics of the loading material had also to be taken into account. There are several types of magnetic distortion to be concerned about. First, there is the production of harmonics as a result of the non-linear magnetization curve of the loading material; second, there is a possible asymmetrical distortion due to hysteresis, and third, there is a possible modulation resulting from the superposition of signals on each other, that is, in effect, a modulation of the head of the wave of one impulse by the tail of the wave of a preceding impulse. The first two of these are effective at the sending end of the cable and the third near the receiving end.

A computation of distortion, including the peculiar magnetic effects, by a steady state a.c. method based on measurements of short loaded conductors indicated that the cable should operate satisfactorily with ordinary sending voltages. Further evidence that none of these various types of distortion would be of serious consequence and that the distortion of a loaded cable could be corrected by terminal apparatus, was obtained by experiments with an artificial line constructed to simulate closely, with regard to electrical characteristics, the type of loaded conductor with which we were then experimenting. This artificial line was loaded with iron dust core coils which served the purpose admirably, not only as regards inductance and alternating current resistance but also as regards magnetic distortion. Iron dust is, of course, very different in its magnetic characteristics from permalloy. However, owing to the large number of turns on a coil, it is operated at much higher field strengths and on a part of the magnetization curve corresponding approximately to that at which permalloy is operated on the cable. The case for magnetic distortion was in fact a little worse with the artificial line than with the then proposed cable. Fig. 5 shows a photograph of the artificial line, the coils of which are in the large iron pots and the resistance and paper condenser capacity units of which are in the steel cases. This line was equivalent to a 1,700 nautical mile cable loaded with 30 millihenries per n.m. and over it legible

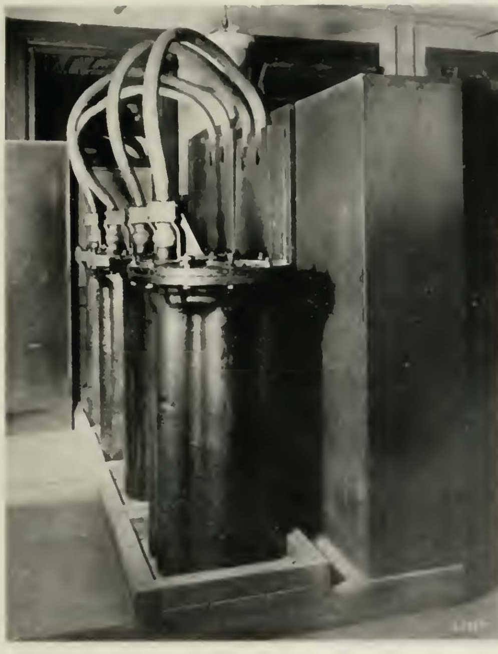


Fig. 5—Loaded Artificial Line

signals were secured at speeds up to more than 2,600 letters per minute. Such a speed of operation was quite beyond the range of the then available telegraph instruments, and accordingly special transmitting and receiving instruments were required. The multiplex distributor of the Western Electric printing telegraph system proved an excellent transmitter for experimental purposes and, for receiving, use was made of a combined vacuum tube amplifier and signal shaping network, the signals being recorded on a string oscillograph. Fig. 6 shows part of a test message received over the loaded artificial cable at a speed of 2,240 letters per minute.

The results of the tests with the artificial loaded cable were entirely in agreement with our calculations and showed that it was

Fig. 6—Test Message. Signals received April 16, 1920, over coil-loaded artificial line equivalent to a 1700 n.m. cable with 30 m.h./n.m. Speed 2240 letters per minute

possible to obtain satisfactory signal shape with a coil-loaded cable having alternating current resistance and distortion factors approximating those of the permalloy-loaded cable. The exact behavior of the proposed cable, including such factors as sea-return resistance and a somewhat variable distributed inductance, could not, of course, be duplicated without prohibitive expense. The approximation was considered, however, to be sufficiently good to justify proceeding with a loaded cable installation so far as questions of signal shaping were concerned. It is interesting to note that the factor which limited the operating speed of the artificial loaded cable was one which is not present in a continuously loaded cable but which would possibly be a serious factor in the operation of a coil loaded cable, namely the oscillations [11] resulting from the finite size and separation of the inductance units.

Operation of Loaded Cables

With the completion of the artificial loaded cable tests there was still one principal question of transmission which had to remain unanswered until a cable had been installed. This was the question of balancing the cable for duplex operation. Ordinary submarine cables are generally operated duplex, the total speed in the two directions being usually from about 1.3 to 2 times the maximum simplex or one-way speed. Except in cases where the external interference is very bad, the limiting speed of duplex operation is determined by the accuracy with which an artificial line can be made the electrical equivalent of the cable. Ordinarily the artificial line is

[11] Carson, Trans. A. I. E. E., Vol. 38, p. 345, 1919.

made up only of units of resistance and capacity arranged to approximate the distributed resistance and capacity of the cable. Sometimes inductance units are added to balance the small inductance which even a non-loaded cable has. In the actual operation of cables, artificial lines are adjusted with the greatest care and a remarkable precision of balance is obtained. This is necessary because of the great difference in current amplitude of the outgoing and incoming signals, the former being of the order of 10,000 times the latter. It is quite obvious that it will be much more difficult to secure duplex operation with a loaded than with an ordinary cable, since not only do the copper resistance and the dielectric capacity have to be balanced, but the artificial line must also be provided with inductance and alternating current resistance. Also the sea-return resistance and inductance which vary with frequency must be balanced.

In view of these difficulties it will probably be impossible to get as great a proportionate gain from duplex operation of loaded cables as is secured with ordinary cables. However, it is quite evident that it will be possible to secure duplex operation at some speed, since, with loaded as with non-loaded cables, the ratio of received to sent current increases rapidly as the speed is reduced and on this account it is much easier to duplex the cable at low speeds than at high. To make duplexing worth while on a cable with approximately equal traffic loads in both directions it is in general only necessary to get a one-way duplex speed half as great as the simplex speed. In fact in some cases the operating advantages of duplex would warrant even a slower duplex speed. On the other hand, there are cables on which the traffic is largely undirectional through most of the day and which would accordingly require a one-way duplex speed somewhat higher than half the simplex speed to justify duplex operation. Whether a sufficiently great speed of duplexing could be secured to justify designing a cable on the basis of duplex operation could not be judged in advance of laying the first cable, and accordingly it was decided to engineer that cable on the basis of simplex operation.

Although it was expected that the new cable might at first have to be operated simplex it should not be supposed that any great difficulty or loss of operating efficiency was anticipated on this account. The speed of the New York-Azores cable is so great that to realize its full commercial advantage practically requires working it on a multi-channel basis as, for example, with a Baudot code, multiplex system, similar to that used on land lines. Such a system may be conveniently adapted to automatic direction reversal and with this modification most of the common objections to simplex operation are removed. Indeed, simplex operation may in this case possess a real advantage over duplex from the commercial point of view since it permits dividing the carrying capacity of the cable most efficiently to handle the excess of traffic in one direction.

Although means have been made available for making efficient use of the loaded cable it should be recognized that the method of operation best suited to satisfy commercial demands must be determined from future experience with cables of the new type. This is especially true with regard to relatively short cables. The discussion of the loaded cable problem in this paper has been confined wholly to the realm of long ocean cables where the limitations of the cable rather than terminal equipment or operating requirements determine the best design. This is the simplest case and the one which at present seems to show the greatest gain from loading. Where traffic requirements are limited and where there is no prospect of ever requiring higher speed than can be obtained with a non-loaded cable of reasonable weight, the advantage of loading is less and becomes smaller as the weight of non-loaded cable which will accomplish the desired result decreases. It should not be concluded, however, that loading will not find important application to short cables. Many short cables are parts of great systems and must be worked in conjunction with long cables. In such cases it may pay to load short sections where otherwise loading would not be justified. Permalloy loading also offers great possibilities for multiple-channel carrier-telegraph operation on both long and short cables and with this type of operation in prospect it is too early, now, to suggest limits to the future applications of permalloy to cables or to predict what will be its ultimate effect on transoceanic communication.

Useful Numerical Constants of Speech and Hearing

By HARVEY FLETCHER

NOTE. The material given in this paper was prepared in a more condensed form for publication in the International Critical Tables. In order to make it available in convenient form for the use of telephone engineers it was deemed advisable to publish it in this journal. The author is indebted to Dr. J. C. Steinberg for able assistance in collecting and arranging the material.

I. Bibliography

A BIBLIOGRAPH of papers on Pitch Discrimination, Intensity Discrimination, Absolute Sensitivity of the Ear, Upper Limit of Audibility, Lower Limit of Audibility, Theories of Hearing and other miscellaneous works on Speech and Hearing are given in a paper by H. Fletcher, *Bell Tech. Jour.*, Vol. II, 4, pp. 178 180, Oct., 1923.

II. Absolute Sensitivity of the Ear

The sensitivity is the minimum audible rms pressure in dynes cm^{-2} in ear canal. The values below are the average of the results of Wien (*Arch. f. ges. Physiol.* 97, p. 1, 1903), Fletcher and Wegel (*Phys. Rev.*, 19, p. 553, June, 1922), and Kranz (*Phys. Rev.*, 21, p. 573, May, 1923) weighted 3, 72, and 14, respectively according to number of ears tested

TABLE I

Frequency (dv)[1]	64	128	256	512	1024	2048	4096
Sensitivity (dynes)	.12	.021	.0039	.001	.00052	.00041	.00042

III. Minimum Audible Power for a Normal Ear

The power in microwatts passing through each square centimeter in the wave front of a free progressive wave in air under average conditions is related to the rms pressure in dynes by the formula

$$p = 20.5\sqrt{J}.$$

The figures of Table I may be converted by this formula to minimum audible powers. It is thus seen that the minimum audible accoustical power is at frequencies between 2,000 and 4,000 vibrations per second and is equal to 4×10^{-10} microwatts per square centimeter

[1] The symbol dv is used to denote "double" or complete vibrations.

IV. Range of Audition in Frequency and Intensity

In Fig. 1 the lower curve is a plot of the average sensitivity values given in Table I. The upper curve gives the pressures that produce a sensation of feeling and serves as a practical limit to the range of auditory sensation. (Wegel, *Bell Tech. Jour.*, 1, p. 56,

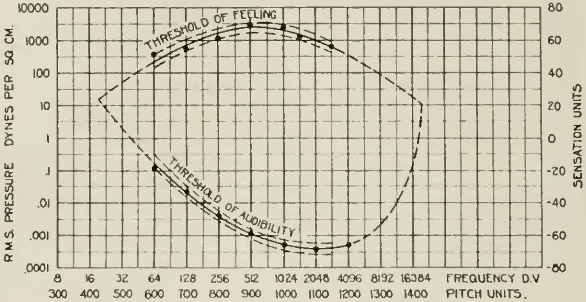


Fig. 1

November, 1922.) Investigators vary from about 8 to 40 dv for the lower pitch limit and from about 12,000 to 35,000 dv for the upper limit. (See 1.) The values of 20 and 20,000 dv shown on the chart were taken as being most representative. Half of the observations lie within the dotted curves. The pitch is equal to 100 $\log_2 N$ and the sensation units equal to 20 log P where N is the frequency and P is the pressure. (Fletcher, *Jour. Frank. Inst.*, 194,

V. Minimum Perceptible Increase in Intensity and Frequency

(Knudsen, *Phys. Rev.* 21, p. 84, Jan., 1923)

Sensation Level in Sensation Units or TU's	Per Cent Increase in Intensity to be Just Perceptible
10	23
20	14
30	12
40	11
50	10 6
60 to 100	10

Frequency	Per Cent Increase in Frequency to be Just Perceptible
64	.93
128	.59
256	.40
512	.32
768 to 4096	.30

p. 289, Sept., 1923.) The sensation level S of a sound is defined by $S=20 \log \frac{P}{P_0}$ where P_0 is the threshold pressure, or it is the number of sensation units above the threshold of audibility. These sensation units are the same as the transmission units used in telephone engineering.

The per cent increase in frequency to be just perceptible varies with sensation level in about the same way as does the per cent increase in intensity to be just perceptible. The values are for monaural reception the tones being heard successively.

VI. The Number of Double Vibrations Necessary to Determine Pitch

(Bode, *Psychol. Stud.*, 2, p. 293, 1907)

TABLE II

	Weak Tones		Medium Tones	
Freq. dv	Time (sec.)	No. of dv	Time (sec.)	No. of dv
128	0.0496	12.1		
256			0.06908	17.6
384	.0672	24.08	0.0445	17.1
512	.0579	29.64	0.04274	21.8

VII. The Masking Effect of One Sound Upon the Audibility of Another Sound

(Wegel and Lane, *Phys. Rev.*, 23, p. 266, Feb., 1924)

If the ear is stimulated by a pure tone of frequency N_1, it is in general rendered less sensitive to other pure tones. The tone that constantly stimulates the ear is called the masking tone. The tone that is heard in the presence of this stimulating tone is called the masked tone. The masking is measured in sensation units or TU's. It is equal to $20\times\log_{10}$ of the ratio of the pressures necessary to perceive the masked tone with and without the presence of the masking tone. In other words it is equal to the number of units that the threshold has been shifted. Fig. 2 shows the amount of masking (ordinate) of tones of various frequencies as a function of the sensa-

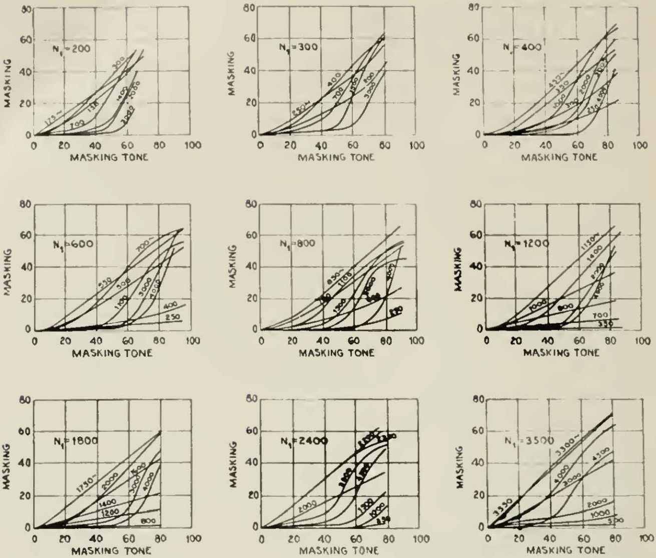


Fig. 2—Masking for Tones in Same Ear

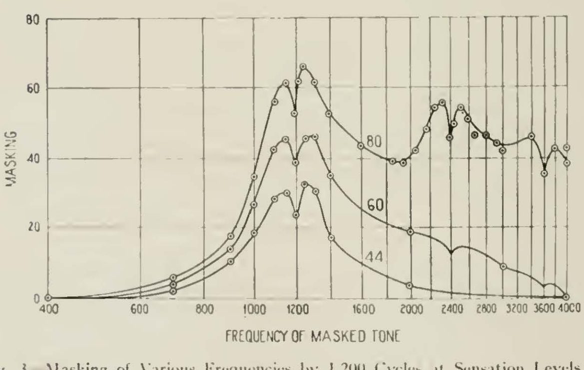


Fig. 3—Masking of Various Frequencies by 1,200 Cycles at Sensation Levels of 80, 60, and 44 Units, Respectively

tion level (abscissa) and frequency N_1 of the masking tone. In Fig. 3 data for a masking tone of 1,200 dv is plotted in which the frequencies of the masked tones are plotted on the abscissa. In order to get satisfactory curves of this kind it is necessary to take more

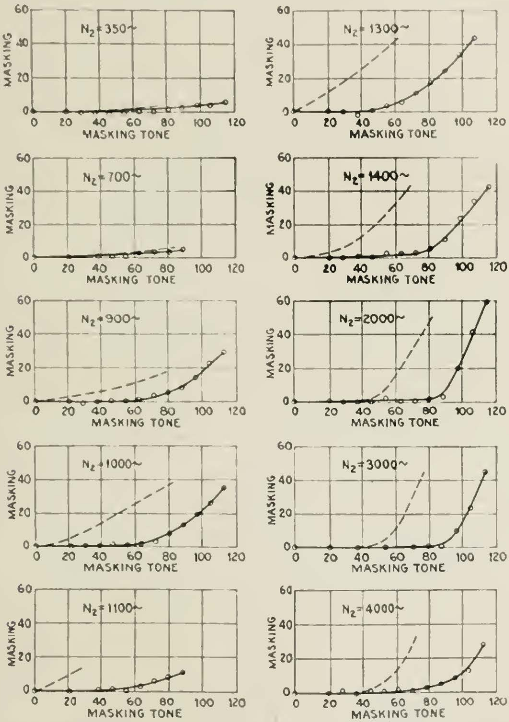


Fig. 4—Masking Data. Tones in Opposite Ears. Masking Tone 1,200 Cycles

comprehensive data than that shown in Fig. 2. The solid curves of Fig. 4 show the masking when the masked and masking tones are introduced into opposite ears. The dotted curves were taken from Fig. 2.

VIII. Conduction of Skull Between the Two Ears

A comparison of the two curves in Fig. 4 shows that the attenuation introduced by the skull from one ear to the other when the tone is introduced by a telephone receiver is between 40 and 50 sensation units corresponding to an intensity ratio of from 10^4 to 10^5. This becomes 7 TU greater when rubber caps are interposed between the head and the receiver cap.

IX. Localization of Pure Tones as a Function of the Phase Difference at the Two Ears

(G. W. Stewart, *Phys. Rev.*, 25, p. 425, May, 1920)

The experimental results can be represented by the formula

$$\frac{\Phi}{\Theta} = 0.0034N + .8 \text{ (approx.)}$$

Φ is the phase difference in degrees of the tones at the two ears.

Θ is the number of degrees to the right or left of the median plane that an observer locates the source of sound. The direction of location is toward the ear leading in phase.

N is the frequency of the tone in dv. The relation applies only for frequencies of 100 to 1,000 dv., inclusive.

X. Constants Used in the Computation of the Loudness of a Complex Sound

(Fletcher and Steinberg, *Phys. Rev.*, 24, p. 306, Sept., 1924)
(Steinberg, *Phys. Rev.* To be published soon)

If L be the loudness as judged by an average normal ear, then

$$L = 3.33 \log_{10}\left[\sum_{n=1}^{n=k} (W_n p_n)^{\frac{2}{r}}\right]^{\frac{r}{2}}$$

where

p_n = rms pressure of the n^{th} component,

W_n = a weight factor for the n^{th} component (Fig. 5)

r = a root factor (Fig. 5)

The sensation levels (See IV) given in the chart are for the complex tone.

XI. Dynamical Constants of the Hearing Mechanism

(Howell, W. H., "A Textbook of Physiology"

(Wrightson, Sir Thomas, "Analytical Mechanism of the Internal Ear")

(*a*) Ear Canal

Length, 2.1–2.6 cm.
Volume, 1 cm^3.
Area at Opening, .33 to .50 cm^2.

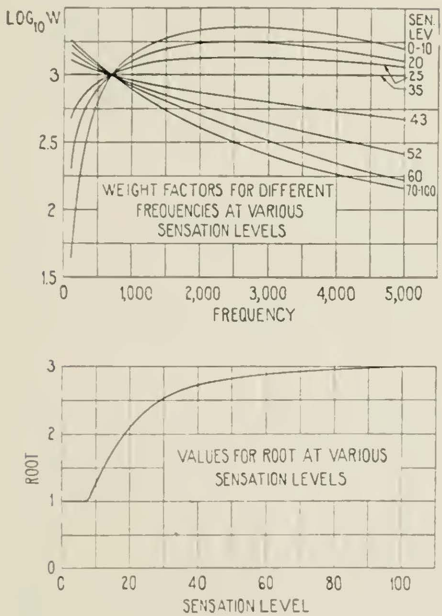


Fig. 5

(*b*) Drum

Vertical Diameter, .85 cm.
Horizontal Diameter, 1.00 cm.
Area, .65 cm^2.

(c) Hammer
Length, .8 to .9 cm.
Weight, 23 mg.

(d) Anvil
Weight, 25 mg.

(e) Stirrup
Weight, 3 mg.

(f) Mechanical Impedance of the Ear Drum
(Data by Wegel and Lane, Bell Telephone Laboratories)

The order of magnitude is 20 to 30 mechanical ohms (cgs units) over the frequency range from 200 to 4,000 dv.

XII. Speech Energy

A. Speech Power

(Data furnished by C. F. Sacia and L. J. Sivian, Bell Telephone Laboratories)

1. The average speech power delivered by an average speaker is about 10 microwatts. In the process of obtaining the average the silent intervals were included. If they are excluded the average increases about 50%. The peak power frequently rises to 2,000 microwatts.

2. Variation of average speech power delivered by different persons during conversation. (Fig. 6.)

B. Energy Frequency Distribution of Average Speech

(Crandall and MacKenzie, *Phys. Rev.*, 19, p. 221, March, 1922) (Fig. 7)

C. Acoustic Power in Vowel Sounds

(Data furnished by C. F. Sacia of the Bell Telephone Laboratories.

This data together with a description of the apparatus and methods used in obtaining it will be given in a paper soon to be published.)

Table III contains data on the power of individual vowels obtained from analyzing the vowel portions of the syllables shown in the key-word. The first two columns give the average power in microwatts of 8 males and 8 females during the particular cycle of the fundamental containing the maximum energy for unaccented vowels. A rough estimate of the corresponding figures of typical accented

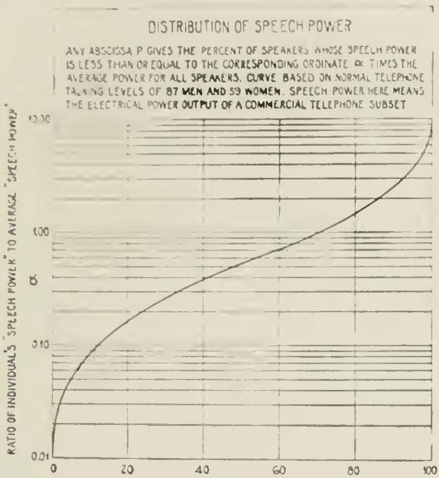


Fig. 6

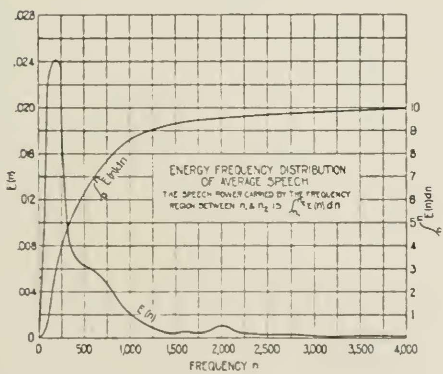


Fig. 7

vowels may be obtained by multiplying these values by a factor of 3. The third and fourth columns give peak factors which convert the power figures of the first two columns into maximum instantaneous powers. Columns 5 and 6 give the maximum values of these peak factors found among the male and female voices, respectively.

TABLE III

Acoustic Power in Microwatts of the Vowel Sounds

Vowel	Key	(1) P_m 8 males	(2) P_m 8 fem.	(3) Av. Peak Factor 8 males	(4) Av. Peak Factor 8 fem.	(5) Max. Peak Factor 8 males	(6) Max. Peak Factor 8 fem.
ū	tool	27	41	2.6	2.8	3.8	3.4
u	took	32	49	4.0	3.1	4.9	3.4
ō	tone	33	44	4.1	3.4	6.4	4.9
o′	talk	37	49	4.5	3.3	5.7	3.6
o	ton	29	38	4.6	3.9	6.8	5.7
a	top	50	48	4.2	3.6	4.2	4.7
a′	tap	43	39	5.4	4.7	7.4	5.2
e	ten	25	30	5.6	3.8	6.3	4.6
ā	tape	21	30	5.3	4.5	6.0	5.1
i	tip	25	31	4.1	3.8	5.8	5.7
ē	team	32	23	4.7	2.6	5.8	3.6

XIII. Frequency of Occurrence of English Speech Sounds

(Table IV contains data from a book by Godfrey Dewey, "The Relative Frequency of English Speech Sounds," Harvard University Press)

TABLE IV

Relative Frequency of Occurrence of English Speech Sounds

Speech Sound	Key	Rel. Freq.	Speech Sound	Key	Rel. Freq.
a	top	3.3	g		0.74
ā	tape	1.84	h		1.81
a′	tap	3.95	j		0.44
e	ten	3.44	k		2.71
ē	eat	2.12	l		3.74
er	term	0.63	m		2.78
i	tip	8.53	n		7.24
ī	dike	1.59	ng	hang	0.96
o	ton	6.33	p		2.04
ō	tone	1.63	r		6.88
o′	talk	1.35	s		4.55
u	took	0.71	sh	shell	0.87
ū	tool	1.89	th	(thin)	37
ou	our	0.59	th	then	3.43
b		1.81	t		7.13
ch	chalk	0.52	v		2.28
d		4.31	w		2.08
f		1.84	y		0.60
			z		2.97

XIV. Interpretation of Speech

(Fletcher, H., *Jour. Frank. Inst.*, 193, 6, June, 1922)

A measure of the interpretation of speech was obtained by means of articulation tests. Meaningless syllables were pronounced and observers were required to record the syllables. The articulation is

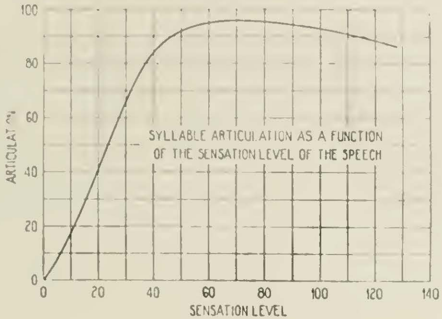


Fig. 8

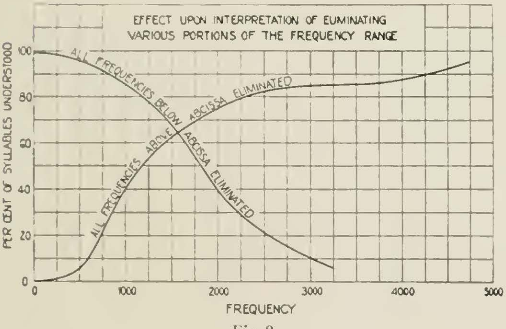


Fig. 9

the per cent of syllables that were correctly recorded. The articulation depends upon the sensation level of the speech (Fig. 8), and upon the width of the frequency band transmitted (Fig. 9).

The syllables that were recorded in these tests were analyzed to show the articulation of the fundamental speech sounds. Fig. 10

shows these articulations as functions of the sensation level of the speech. In Fig. 11 they are shown as functions of the width of the transmitted frequency band. It should be noted that the term articulation as here employed denotes only the correct interpretation of unrelated speech sounds and is not a measure of voice naturalness which is also an important factor in the telephonic transmission of speech.

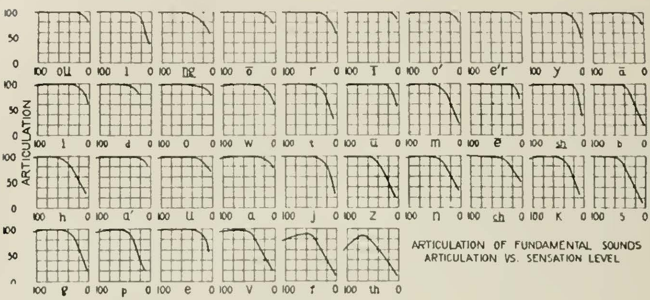


Fig. 10

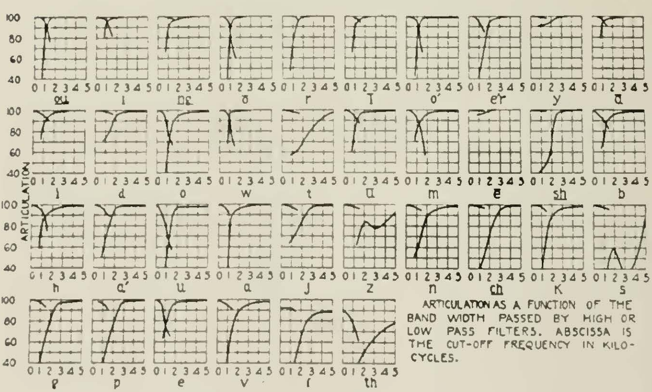


Fig. 11

Graphic Representation of the Impedance of Networks Containing Resistances and Two Reactances

By CHARLES W. CARTER, Jr.

ABSTRACT: The driving-point impedance of an electrical network composed of any number of resistances, arranged in any way, and two pure reactances, of any degree of complication within themselves but not related to each other by mutual reactance, inserted at any two points in the resistance network, is limited to an eccentric annular region in the complex plane which is determined by the resistance network alone.

The boundaries of this region are non-intersecting circles centered on the axis of reals. The diameter of the exterior boundary extends from the value of the impedance when both reactances are short-circuited to its value when both are open-circuited. The diameter of the interior boundary extends from the value of the impedance when one reactance is short-circuited and the other open-circuited to its value when the first reactance is open-circuited and the second short-circuited.

When either reactance is fixed and the other varies over its complete range, the locus of the driving-point impedance is a circle tangent to both boundaries. By means of this grid of intersecting circles the locus of the driving-point impedance may be shown over any frequency range or over any variation of elements of the reactances. This is most conveniently done on a doubly-sheeted surface.

The paper is illustrated by numerical examples.

Introduction

SUPPOSE that any number of resistances are combined into a network of any sort and provided with three pairs of terminals, numbered (1) to (3) as in Fig. 1. The problem set in this paper is to investigate the driving-point impedance[1] of such a network at

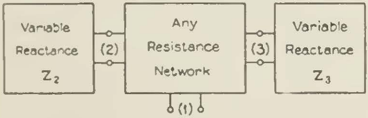


Fig. 1—The Network to be Discussed

terminals (1) when variable pure reactances, Z_2 and Z_3, are connected to terminals (2) and (3), respectively. Z_2 and Z_3 are formed of capacities, self and mutual inductances. They are not connected to each other by mutual reactance, but they may be of any degree of complication within themselves.

The problem is dealt with in terms of the complex plane: that is, the resistance components of the impedance, S, measured at terminals

[1] 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.

(1) are plotted as abscissas and the reactance components as ordinates. To every value of the impedance, then, there is a corresponding point, and to the values of the impedance over a range of variation of some element, or over a frequency range, there corresponds a locus, in the complex plane. This locus may be labelled at suitable points with the corresponding value of the variable. So labelled, it combines into one the curves which are usually plotted to show separately the variation of the reactance and resistance components or to show separately the variation of absolute value and angle.

The use of the complex plane is not new: it is the basis of most of the vector diagrams for electrical machinery. The characteristics of both smooth and loaded transmission lines have also been displayed by its means. Its application to electrical networks, however, is not common, and it is a subsidiary purpose of this paper to illustrate the fact that the properties of certain networks, which have complicated characteristics if exhibited in the usual way, may be shown quite simply in the complex plane. This simplicity, combined with generality, is attained by application of theorems concerning functions of a complex variable which are immediately available.

The Fundamental Equations

The impedance measured in branch 1 of any network is

$$S=R+iX=\frac{\Delta}{\Delta_{11}} \tag{1}$$

where Δ is the discriminant of the network, either in terms of branches or n independent meshes.[2]

Assigning the reactances Z_2 and Z_3 to meshes 2 and 3

$$\Delta=\begin{vmatrix} R_{11} & R_{12} & R_{13} & . & . & R_{1n} \\ R_{21} & R_{22}+Z_2 & R_{23} & . & . & R_{2n} \\ R_{31} & R_{32} & R_{33}+Z_3 & . & . & R_{3n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ R_{n1} & R_{n2} & R_{n3} & . & . & R_{nn} \end{vmatrix} \tag{2}$$

where R_{jj} is the resistance in mesh j and $R_{jk}(=R_{kj})$ that common to meshes j and k.

[2] See: G. A. Campbell, Transactions of the A. I. E. E., 30, 1911, pages 873–909, for a complete discussion of the solution of networks by means of determinants.

Therefore
$$S=\frac{\Delta+\Delta_{22}Z_2+\Delta_{33}Z_3+\Delta_{22\cdot 33}Z_2Z_3}{\Delta_{11}+\Delta_{11\cdot 22}Z_2+\Delta_{11\cdot 33}Z_3+\Delta_{11\cdot 22\cdot 33}Z_2Z_3}\tag{3}$$

where Δ is the discriminant of the resistance network alone and $\Delta_{jj\cdot kk\cdot ll}$ denotes the cofactor of the product of the elements of Δ located at the intersections of rows j, k and l with columns j, k and l, respectively.

For convenience this is written as

$$S=\frac{a+bZ_2+cZ_3+dZ_2Z_3}{a_1+b_1Z_2+c_1Z_3+d_1Z_2Z_3}.\tag{4}$$

The constants of (3) and (4) are real and positive since they are cofactors of terms in the leading diagonal of the discriminant of a resistance network. The determinant being symmetrical, there is the following relation among them:

$$(ad_1-a_1d+bc_1-b_1c)^2=4(bd_1-b_1d)(ac_1-a_1c).\tag{5}$$

The function to be studied is, then, a rational function of two variables, having positive real coefficients determined by the resistances alone. Furthermore, if one reactance is kept constant while the other is varied, the function is bilinear. The particular property of the bilinear function, which has been studied in great detail, of interest here, is that by it circles are transformed into circles.[3]

When, as in this case, the variable in a bilinear function is a pure imaginary, the function may be rewritten in a form which gives directly the analytical data needed. For suppose

$$w=\frac{u+vz}{u_1+v_1z}\tag{6}$$

where z is a pure imaginary and the coefficients are complex. This is

$$w=\frac{v}{v_1}+\frac{u-u_1v/v_1}{u_1+v_1z}.\tag{7}$$

Multiplying the second term by a factor identically unity,

$$w=\frac{v}{v_1}+\frac{u-u_1v/v_1}{u_1+v_1z}\times\frac{v_1'(u_1+v_1z)+v_1(u_1'+v_1'z')}{u_1v_1'+u_1'v_1}\tag{8}$$

where primes indicate conjugates, or

$$w=\frac{uv_1'+u_1'v}{u_1v_1'+u_1'v_1}+\frac{uv_1-u_1v}{u_1v_1'+u_1'v_1}\times\left(\frac{u_1'+v_1'z'}{u_1+v_1z}\right).\tag{9}$$

[3] G. A. Campbell discusses, in the paper cited, the theorem that if a single element of any network be made to traverse any circle whatsoever, the driving-point impedance of the network will also describe a circle.

Now, as z is varied, the first term is constant. In the second term the first factor is constant and the second factor varies only in angle, since the numerator is the conjugate of the denominator. The first term, therefore, is the center, and the absolute value of the first factor of the second term is the radius, of the circle in which w moves as z takes all imaginary values.

One Variable Reactance Giving Circular Locus

The significance of the equations may be made apparent by a study of Fig. 2, which shows the impedance S when one of the reactances, say Z_3, is made zero. We have, then,

$$S=\frac{A+A_{22}Z_2}{A_{11}+A_{11\cdot 22}Z_2}=\frac{a+bZ_2}{a_1+b_1Z_2} \tag{10}$$

and the trivial case $ab_1-a_1b=0$ is excluded. This is of the type of (6). When Z_2 varies over all pure imaginary values, S traces out a

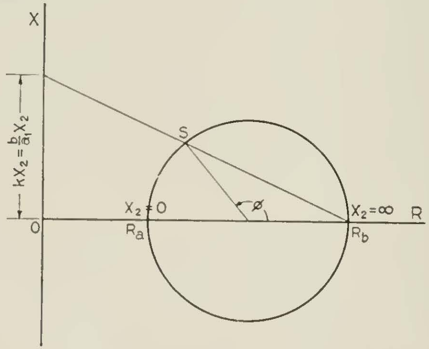


Fig. 2—Locus of the Impedance S with One Variable Reactance

circle, which (9) shows has its center on the resistance axis. Its intercepts on the resistance axis are

$$S=\frac{a}{a_1}=R_a,\text{ say, when } Z_2=0 \tag{11}$$

and

$$S=\frac{b}{b_1}=R_b,\text{ when } Z_2=\infty. \tag{12}$$

But in a symmetrical determinant

$$A_{11}A_{22} - A_{12}^2 = A A_{11\cdot 22}; \tag{13}$$

therefore

$$ab_1 < a_1 b \tag{14}$$

or

$$\frac{a}{a_1} < \frac{b}{b_1} \tag{15}$$

whence

$$R_a < R_b. \tag{16}$$

To find the value of S when Z_2 has some value, say $Z_2 = iX_2$, it is only necessary to mark the circular locus with a scale in terms of Z_2. This may be done directly by using (9) to determine the angle, ϕ, which the radius of the circle makes when $Z_2 = iX_2$. It is simpler to use the fact that a line passing through R_b and the point S has an intercept on the reactance axis of

$$X_0 = \kappa X_2 \tag{17}$$

where $\kappa = b/a_1$.

The factor κ is determined by the resistances; therefore the scale, as well as the locus, is completely fixed by the resistances. Since κ is always positive, as X_2 is increased the circle is traversed in a clockwise sense; for positive values of X_2 the upper semi-circle is covered; for negative values, the lower. That is, when Z_2 is an inductance the impedance of the network varies on the upper semi-circle from R_a to R_b as the frequency is increased from zero to infinity. When the magnitude of Z_2 is changed the same semi-circle is described but each point (except the initial and final ones) is reached at a different frequency. When Z_2 is a capacity the lower semi-circle, from R_b to R_a, is traced out.

We know that, in general, the value of a pure reactance [4] increases algebraically with frequency, and that its resonant and anti-resonant frequencies alternate, beginning with one or the other at zero frequency. When Z_2 is a general reactance, therefore, as the frequency increases the entire circle is described in a clockwise sense between each consecutive pair of resonant (or anti-resonant) frequencies. For example, if Z_2 is made up of n branches in parallel, one being an inductance, one a capacity and the others inductance in series with capacity, as the frequency increases from zero to infinity the circle is traced out completely $n-1$ times commencing with R_a.

[4] See: A Reactance Theorem, R. M. Foster, *Bell System Technical Journal*, April, 1924, pages 259–267; also: Theory and Design of Uniform and Composite Electric Wave-Filters, O. J. Zobel, *Bell System Technical Journal*, January, 1923, pages 1–47, especially pages 35–37.

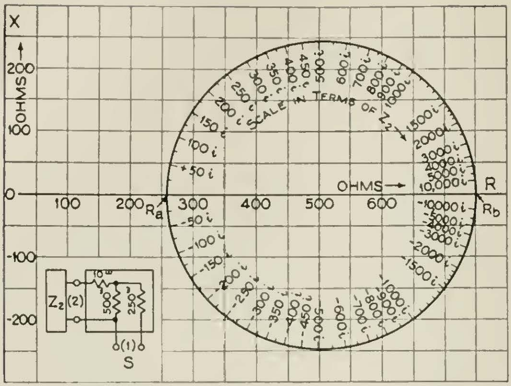


Fig. 3—Impedance of Resistance Network Containing One Variable Reactance

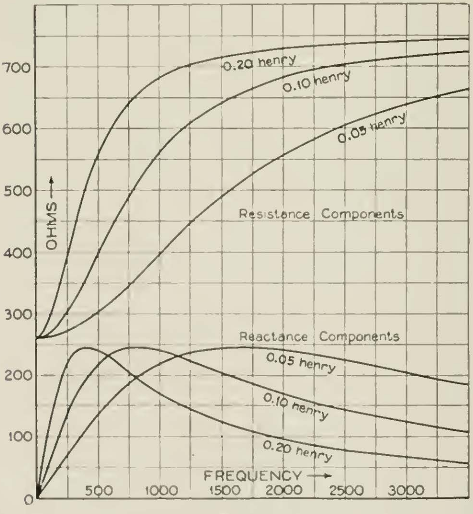


Fig. 3a Components of Impedance in Fig. 3 when Z_2 is an Inductance Having the Values 0.05, 0.10, and 0.20 Henry

In Fig. 3 is shown the impedance locus for the particular network given on the diagram. The circle is marked in terms of Z_2. From it, certain properties of S may be read at once: the resistance component, R, varies between 260 and 750 ohms, and the reactance

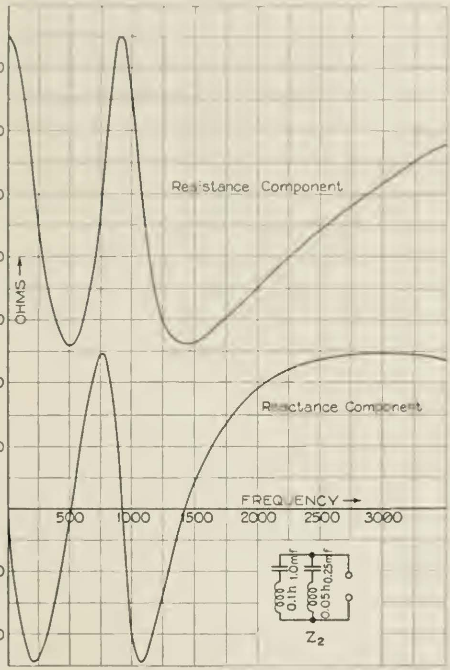


Fig. 3b—Components of Impedance in Fig. 3 when Z_2 is Doubly-Resonant

component, X, is not greater than 245 ohms nor less than -245 ohms, attaining these values when Z_2 is $+510i$ and $-510i$, respectively.

When the variation of the reactance Z_2 with frequency is known the variation of R and X with frequency may be found by using the scale on the circle. For a particular reactance network, the scale may be marked directly in terms of frequency, or if it is desired to compare the behavior of R and X when different reactance networks are substituted, the impedance locus may be marked with the frequency scale for each reactance network in some distinctive manner.

However, to show in the usual way some of the types of R and X curves represented by the locus of Fig. 3, as well as to avoid needless complication of what is intended as an illustrative rather than a working drawing, Figs. 3a and 3b have been prepared by direct projection from Fig. 3. In Fig. 3a are shown the R and X curves plotted against frequency when Z_2 is an inductance. In Fig. 3b are shown similar curves when Z_2 is a doubly-resonant reactance. The R component has a minimum at each resonant frequency and a maximum at each anti-resonant frequency, while the X component becomes zero at resonant and anti-resonant frequencies alike. The number of examples from this one resistance network might be multiplied endlessly; it is believed, however, that these are sufficient to show the great amount of information to be obtained in very compact form from one simple figure in the complex plane, and the especial superiority of the complex plane in displaying the characteristic common to all the curves of Figs. 3a and 3b: namely, that R and X at any frequency, with any reactance network, are such that the impedance lies on one circle.

Two Variable Reactances Giving Eccentric Annular Domain

Returning to the more general impedance of (4) it is seen that in each case short-circuiting and open-circuiting the terminals (2) and (3) one at a time, and varying the reactance across the other terminals, yields a locus for S which is a circle of the type just discussed. These circles are determined as follows:

Circle	Extremities of Diameter	Scale Factor κ
$Z_2=0$	R_a and R_c	c/a_1
$Z_2=\infty$	R_b and R_d	d/b_1
$Z_3=0$	R_a and R_b	b/a_1
$Z_3=\infty$	R_c and R_d	d/c_1

where $R_c=c/c_1$ and $R_d=d/d_1$. An examination similar to that in (13)–(15) shows that

$$R_a \leq R_b \leq R_d, \tag{18}$$

$$R_a \leq R_c \leq R_d. \tag{19}$$

It may furthermore be assumed without loss of generality, since it is merely a matter of labelling the reactances Z_2 and Z_3, that $R_b \leq R_c$.

Hence, the four critical points of the impedance are always in the following order:

$$R_a \leq R_b \leq R_c \leq R_d. \tag{20}$$

These circles are shown in Fig. 4. By means of the appropriate scale factors κ each may be marked in terms of the reactance which is left in the circuit.

Now suppose Z_3 is kept constant at some value which is a pure imaginary, and Z_2 is varied over the range $-i\infty \leq Z_2 \leq +i\infty$. We may rewrite (4) in the normal form (6):

$$S = \frac{a + cZ_3 + (b + dZ_3)Z_2}{a_1 + c_1Z_3 + (b_1 + d_1Z_3)Z_2}. \tag{21}$$

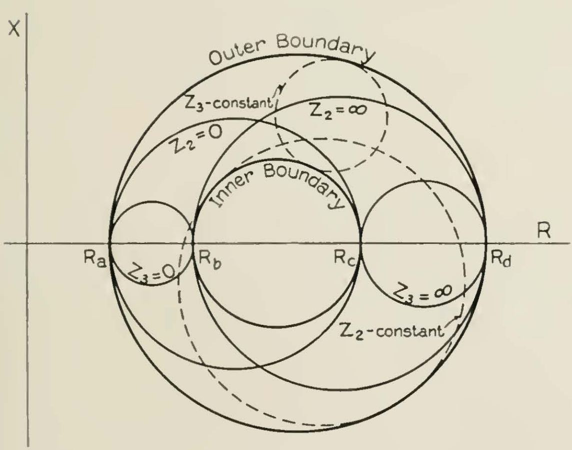


Fig. 4 The Region to which S is Restricted and the Critical Circles

The locus of S is one of a family of circles, each circle corresponding to a value of Z_3 and completely traced out by complete variation of Z_2. The properties of each circle may be found by substitution in (9).

Similarly, if Z_2 is held constant while Z_3 varies, the locus of S is one of another family of circles.

By the use of (9), keeping (5) in mind, it may be shown that the circles of each of these families are tangent to two circles determined by the resistance network alone. Both families are tangent internally to a circle centered on the resistance axis, extending from R_a to R_d. Both are tangent to a circle centered on the resistance axis, extending from R_b to R_c, in such a way that the Z_3-constant circles are tangent *externally* and the Z_2-constant circles are tangent *enclosing* the circle from R_b to R_c. These relationships are illustrated in Fig. 4.

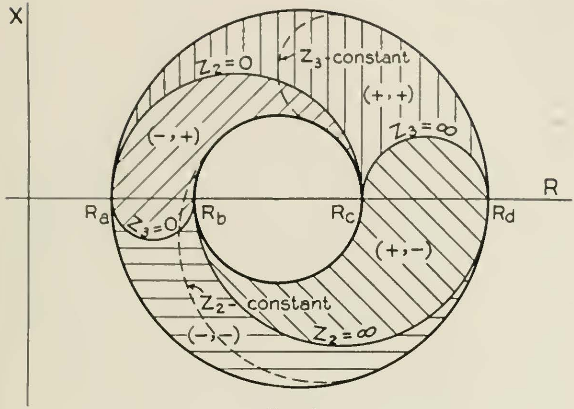
The circles R_a to R_d and R_b to R_c are, therefore, outer and inner boundaries, respectively, of the region mapped out by the two families of circles generated when first one and then the other reactance is treated as a parameter while the remaining reactance is treated as the variable. No matter what reactances may be attached to terminals (2) and (3), the resistance component R, measured at terminals (1), is not greater than the resistance when terminals (2) and (3) are open and not less than the resistance when terminals (2) and (3) are short-circuited, and the reactance component X, measured at terminals (1), is not greater in absolute value than half the difference of the resistances measured when terminals (2) and (3) are open and short-circuited. That is,

$$R_a \leq R \leq R_d, \tag{22}$$

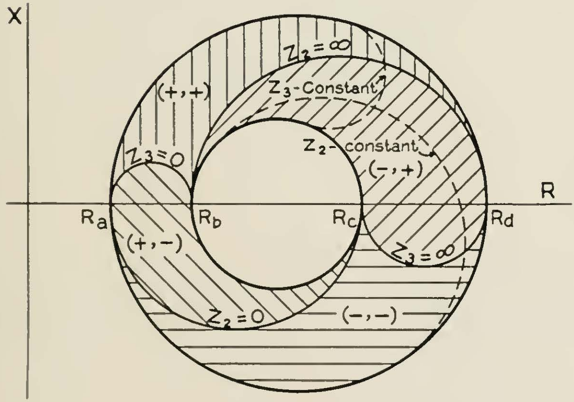
$$|X| \leq \tfrac{1}{2}(R_d - R_a). \tag{23}$$

The two families of circles (Z_2-constant and Z_3-constant) intersect and may be used as a coordinate system from which the components of S may be read for any pair of values Z_2, Z_3. To avoid intersections giving extraneous values of S resort is made to a doubly-sheeted surface, analogous to a Riemann surface, for which the two boundary circles are junction lines. That is, the impedance plane is conceived of as two superposed sheets, transition from one to the other being made at the boundary circles. Thus, in Fig. 5, where the two sheets are separated, each Z_2-constant circle is shown running from the outer to the inner boundary in Sheet I (using the clockwise sense), and from the inner to the outer boundary in Sheet II, while the Z_3-constant circles run from the inner to the outer boundary in Sheet I and are completed in Sheet II.[5]

[5] It may be mentioned that the inner and outer boundaries are impedance curves traced out when $Z_2 Z_3 = \frac{\Delta_{12}\Delta_{13}}{\Delta_{12\cdot 33}\Delta_{13\cdot 22}}$ and $\frac{Z_2}{Z_3} = \frac{\Delta_{13}\Delta_{12\cdot 33}}{\Delta_{12}\Delta_{13\cdot 22}}$, respectively.



Sheet I



Sheet II

Fig. 5—The Doubly-Sheeted Surface

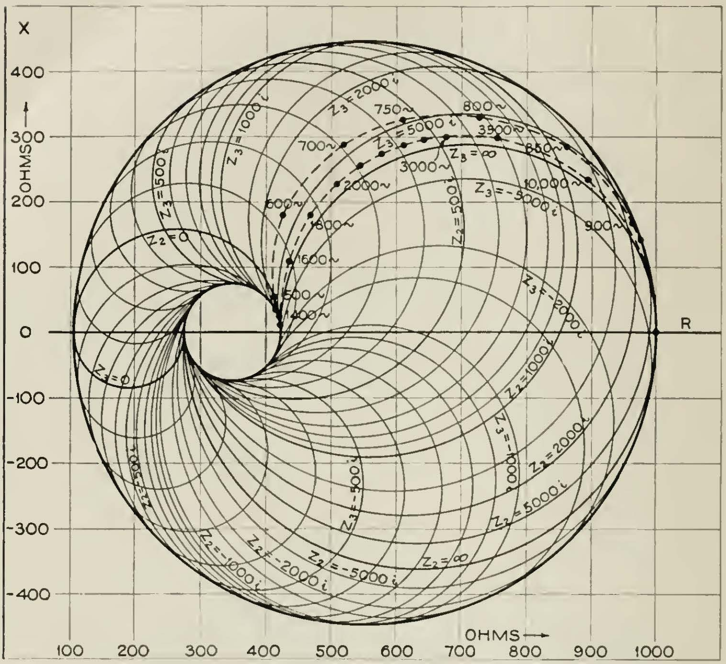


Fig. 6—Sheet I

Sheets I and II of Fig. 6, taken together, show the impedance domain of the network at the bottom of the opposite page, made up of three fixed resistances and two variable reactances. The dashed curve, appearing in four distinct parts, two on each sheet, shows the impedance S when Z_2 is the doubly-resonant circuit of Fig. 3b, and Z_3 is an inductance of 1.0 henry. Points on this curve are labelled in terms of frequency

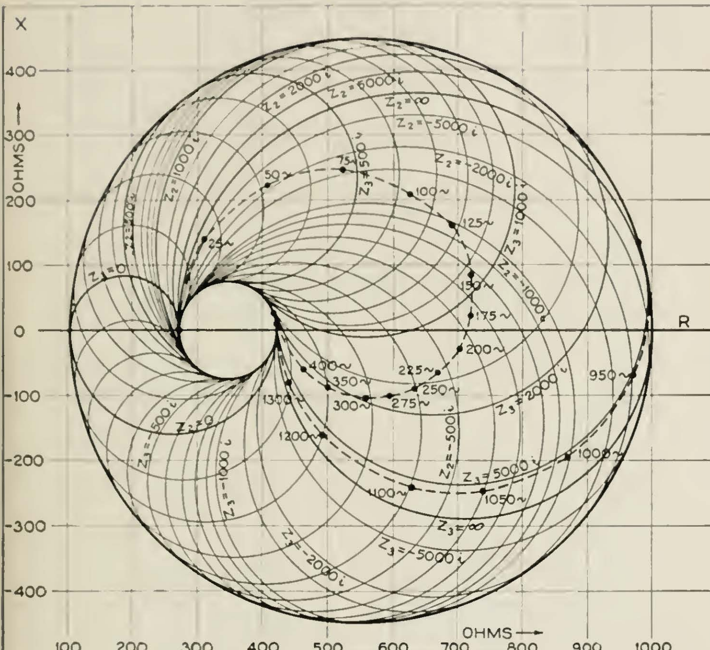
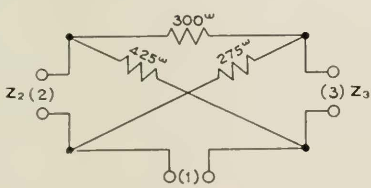


Fig. 6—Sheet II



The numbering of the sheets is, of course, arbitrary. If the upper half of the $Z_2 = 0$ circle is put on Sheet I, the arcs of the other critical circles are determined as follows:[6]

Circle	On Sheet I	On Sheet II
$Z_2 = 0$	Upper half	Lower half
$Z_2 = \infty$	Lower half	Upper half
$Z_3 = 0$	Lower half	Upper half
$Z_3 = \infty$	Upper half	Lower half

Each sheet, then, is divided into four sub-regions, indicated on Fig. 5 by the signs of the reactances for which S is within them. When Z_2 and Z_3 are composed of single elements the sub-regions in which S falls at any frequency are as follows:

Z_2	Z_3	Sub-Region	At Frequency Zero	At Frequency Infinity
Inductance	Inductance	$(+,+)$	$S = R_a$	$S = R_d$
Inductance	Capacity	$(+,-)$	R_c	R_b
Capacity	Inductance	$(-,+)$	R_b	R_c
Capacity	Capacity	$(-,-)$	R_d	R_a

The course of S over the complete frequency range may be shown by a curve through the appropriate intersections of the Z_2-constant and Z_3-constant circles, as in the following example.

The impedance region for a particular bridge network is illustrated in the two sheets of Fig. 6. The arcs of Z_2-constant and Z_3-constant circles in each sheet form a curvilinear grid superposed on the R,X grid of the complex plane. For example, if $Z_2 = 200i$ and $Z_3 = 900i$, the value of S is read from Sheet I as $327 + i291$, and S has this value irrespective of the structure of Z_2 and Z_3.

An impedance curve (dashed) is shown in Fig. 6 representing the variation of S with frequency when Z_2 is the doubly-resonant reactance

[6] When the sheets are numbered in this way, the point S falls on Sheet I or Sheet II according to the following table, in which k_1 and k_2 are the critical values for the product and quotient of Z_2 and Z_3, respectively, given in Footnote 5:

(Z_2, Z_3)	$(+,+)$	$(+,-)$	$(-,+)$	$(-,-)$
On Sheet I, if	$Z_2/Z_3 < k_2$	$Z_2Z_3 > k_1$	$Z_2Z_3 < k_1$	$Z_2/Z_3 > k_2$
On Sheet II, if	$Z_2/Z_3 > k_2$	$Z_2Z_3 < k_1$	$Z_2Z_3 > k_1$	$Z_2/Z_3 < k_2$

For the network of Fig. 6, $k_1 = 116{,}875$ and $k_2 = 0.972111$.

used in Fig. 3b and Z_3 is an inductance of 1.0 henry. This impedance curve has four parts, two in each sheet. It starts on the resistance axis at the intersection of the $Z_2 = \infty$ and $Z_3 = 0$ circles. As the frequency increases from zero the first part of the curve is traced out in Sheet II. At 25 cycles the impedance is approximately $310+i140$. The reactance component has a maximum of about 250 ohms at about 70 cycles, the resistance component has a maximum of about 720 ohms at about 160 cycles, the reactance component has a minimum of about -110 ohms at about 300 cycles, and finally at about 480 cycles the curve reaches the inner boundary, whereupon it changes to Sheet I. It remains in Sheet I up to a frequency of about 910 cycles, the resistance component having a minimum and the reactance component a maximum, which may be read from the diagram. The impedance between 910 cycles and approximately 1,390 cycles lies on Sheet II, and from 1,390 cycles to infinite frequency on Sheet I. The resistance component has a total of three maxima and three minima, and the reactance component three maxima and two minima, following the cyclical order: R-minimum, X-maximum, R-maximum, X-minimum.

An interesting exercise is to observe the effect on the impedance curve of changing the value of the inductance Z_3. The curve intersects the Z_2-constant circles at the same frequencies in each case, but the points of intersection are moved in a clockwise or counterclockwise sense as Z_3 is increased or decreased. With each such change parts of the impedance curve disappear from one sheet and reappear on the other. For instance, with a decrease of the inductance Z_3 the first loop of the impedance curve on Sheet II shrinks, and with sufficient decrease in inductance may become too small to plot, although it does not disappear entirely.

It is evident that if Z_2 and Z_3 are formed of reactance networks of greater complication the impedance curve may be very involved. But no matter how tortuous its path, it is restricted to the impedance region, that is, to the ring-shaped region between the non-intersecting boundary circles determined by the resistance network alone.

My thanks are due to Dr. George A. Campbell for his stimulating, continued interest, and to Mr. R. M. Foster for suggestions on every phase of this work.

The Vibratory Characteristics and Impedance of Telephone Receivers at Low Power Inputs

By A. S. CURTIS

THE ordinary telephone receiver is one of the most sensitive known detectors of weak alternating currents over a considerable part of the audible frequency range. Its high sensitivity, combined with its simplicity and convenience, have led to its general adoption as the detecting element in the AC impedance bridge and other measuring apparatus employing the nul method. There are also a number of cases outside of the laboratory where a knowledge of the behavior of the receiver operating near its minimum audible power input is of importance. In apparatus developed during the World War, such as that for detecting and locating submarines, in radio reception, and in the reception of various other sorts of signals, the receiver is frequently operated near the threshold of audibility. While it is in general possible to employ a vacuum tube amplifier to render weak signals more easily audible, considerations of cost or increased complication often make it impracticable to do so. In any case, if it is desired to reduce to the limit the minimum audible signal, it is necessary to know the constants of the receiver working on these low power inputs, in order to design intelligently its circuits and other associated apparatus.

Current literature dealing with the sensitivity of telephone receivers indicates that the relation between the impedance and vibratory characteristics of the receiver at currents near minimum audibility to those as ordinarily determined in the laboratory, is not generally known. It would, therefore, seem of interest to publish the results of an experimental determination of receiver characteristics at very low currents. Such an investigation was carried on in 1918 and 1919, using the Western Electric No. 509 radio receiver (the present standard Western Electric Receiver for radio use). The work, however, was done, not merely with the idea of determining the characteristics of this particular instrument, but for the purpose of ascertaining the behavior of receivers in general, near minimum audibility.

Inasmuch as the damped impedance of the receiver—that is the impedance with the diaphragm held motionless—is very close to the impedance obtained with the instrument on the ear, it is commonly used as the basis of circuit calculations. A knowledge of its value for weak currents is therefore of importance. Measurements were first made of the damped impedance of six instruments at a frequency of 1,000 cycles for a wide range of input current, and later the work was extended to the measurement of the vibratory characteristics. A bridge network was used for measuring the current supplied to the impedance bridge and from the circuit constants the current through the receiver under test could be calculated. The re-

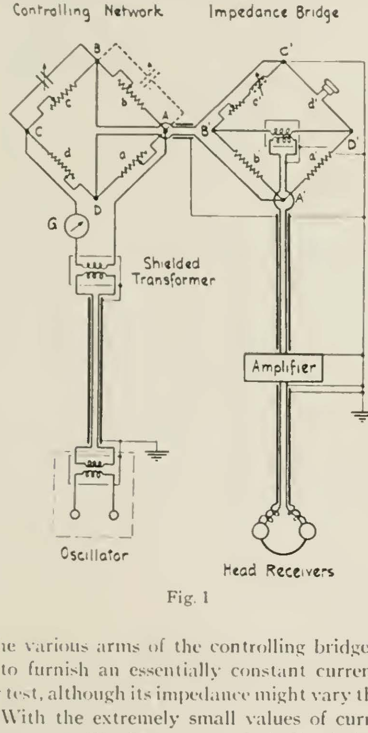


Fig. 1

sistances in the various arms of the controlling bridge network were chosen so as to furnish an essentially constant current through the receiver under test, although its impedance might vary through a rather wide range. With the extremely small values of currents involved, it was necessary to amplify the power to the bridge balancing receivers approximately 100 TU. For this amount of amplification, it was obviously necessary to take extreme precautions in grounding and shielding the apparatus, in order to reduce to inaudibility the effect of stray fields from the source of current supply. This was successfully done and the impedance bridge measured impedance accurately with currents as low as 10^{-9} amperes, through the receiver under test. The correctness of the point of balance of the bridge was established by measurements of standard impedances over the range of currents employed in the receiver tests. A schematic diagram of the circuit is shown in Fig. 1.

For measurements of damped impedance, the receiver was placed in a small sound-proof box, with its diaphragm damped by a micrometer depth gauge, which was carefully adjusted so as just to impinge upon the diaphragm. It was necessary to insulate the receiver from mechanical agitation, since minute voltages generated in it were sufficiently amplified to cause an excessive noise in the head receivers.

Fig. 2 shows the damped effective resistance and reactance of the six instruments, taken at 1,000 cycles, plotted on semi-logarithmic

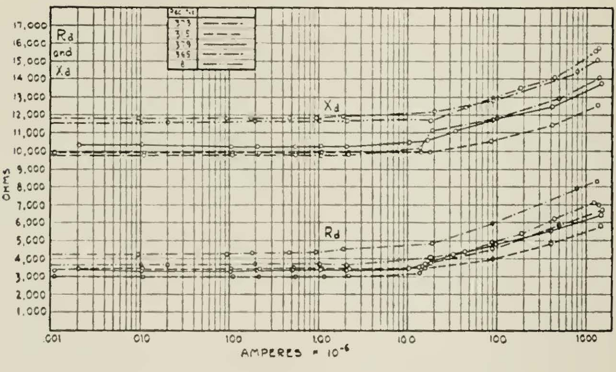


Fig. 2

paper. It will be seen that below approximately 10^{-6} amperes, the impedance is constant. However, above this value both the effective resistance and the reactance show a consistent increase with the current. The minimum current employed (10^{-9} amperes), is between two and three times the minimum audible current for this type of instrument, but from the data taken there is no reason to suppose that the impedance would vary for smaller currents. This receiver has a winding of 11,000 turns, and it can, therefore, be assumed that this type of structure will have constant impedance below a magnetomotive force of .01 ampere turns. For laboratory measurements on this instrument a current of 2×10^{-6} amperes is ordinarily used, and it will be noted that the impedance at extremely low currents is not greatly different.

It is generally known that, in the case of either a steady or an alternating field, the permeability and the shape of the hysteresis

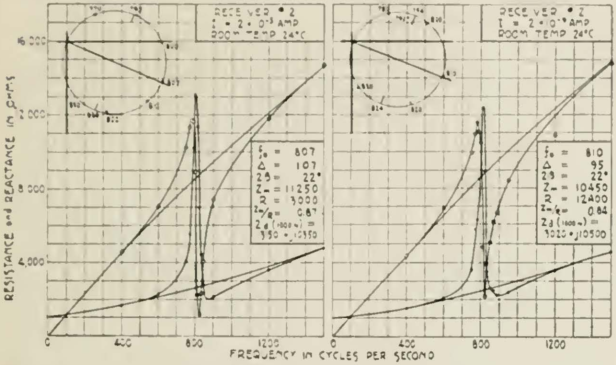


Fig. 3—f_0 = natural frequency; Δ = logarithmic decrement per second; 2β = depression angle of principal diameter; Z_m = maximum motional impedance; R = free resistance at resonance; Z_d = damped impedance.

loop for ordinary magnetic materials reach limiting values as the magnetomotive force is reduced, that is, further reductions of the magnetomotive force have no effect on these magnetic characteristics. The results cited above show that this condition obtains for a weak alternating field when it is superimposed on a relatively strong steady field.

In the measurements of free impedance for determining the vibratory characteristics the small sound-proof box could not be used on account of the proximity of its walls. Accordingly, the receiver and the impedance bridge were placed in a large sound-proof booth with padded walls where the effect of reflection of sound waves was very small. With the diaphragm of the instrument free to vibrate, its efficiency as a sound detector was materially increased and the noise in the head receivers due to the slightest movements of the observer became so serious that it was not feasible to take data with currents of less than 2×10^{-9} amperes.

Fig. 3 shows impedance characteristics, with their associated circles, of the same receiver with currents of 2×10^{-9} and 2×10^{-5} amperes. It will be seen that the differences between these curves are insignificant when one considers the low precision of motional impedance data in the absence of extreme precautions with regard to constancy of temperature, etc. Moreover, other impedance analyses at intermediate values of current agree with the above within the precision of the measurements.

To summarize the results, it may be said that the characteristics of receivers remain substantially unaltered as the current is reduced to the point of minimum audibility. In taking impedance measurements, it is well to use a current which is low enough to be on the flat part of the curve. This can usually be done without the use of amplifiers between the impedance bridge and balancing receivers. The fact that the vibratory characteristics of the receiver remain unaltered as the power input is reduced to the threshold of audibility throws an interesting light on the behavior of the diaphragm material under very small motions. Calculations of the minimum audible amplitude near resonance, based on the fact that the constants of the material remain unchanged, show it to be of the order of 10^{-9} centimeters. This motion is less than the mean molecular diameter of the diaphragm material.

Some Contemporary Advances in Physics—VIII
The Atom-Model, First Part[1]

By KARL K. DARROW

A. Introductory Remarks About Atom-Models

MORE than any other word of the language, the word *atom* is implicated with the history of human speculations concerning the nature of things. It is introduced when people cease to content themselves with observing, and begin to philosophize. There are many of the fundamental and essential writings of the literature of physics in which it does not appear, or appears without warrant. These are the descriptions of things observed, the accounts of experiments, the records of measurements, on which the edifice of theoretical physics is founded. There are many articles of what is commonly called the "theoretical" sort in which it does not occur. Such are the papers on the motions of planets, on the vibrations of elastic solids, on the currents in electrical networks, on the courses of light-rays through optical systems—papers which are essentially descriptions, although they give the impression of being something greater and deeper because they relate to idealized cases, and are phrased in the laconic language of mathematics. When the word *atom* appears justifiably in a discourse, it means that the author has departed from the safe routine of describing observed and observable events, however selectively, however skilfully, however intelligently. It signifies that he has gone beyond the limits of observation, and has entered upon the audacious adventure of constructing by the side of the real universe an ideal one, which shall act as the real one does, and be intelligible through and through.

Atoms are the building stones of this art-world or image-world, which is intended to represent the actual world, imperfectly indeed for the time being, perhaps completely at some distant day. Some few experiments, it is true, prove (as well as anything can prove anything else) the existence of very minute particles of matter having the minute charges, the minute masses, the minute magnetic moments

[1] This part, the first of two composing the article, is devoted chiefly to the facts of observation which the favorite atom-model of the physicists of today—the atom-model known by the names of Rutherford and of Bohr—is designed to interpret. A brief description of this atom-model is included; but the detailed account of the peculiar features, of the strange and important limitations which are imposed upon it to adjust it to all the phenomena mentioned, is reserved for the second part. Owing to the great quantity of information which it is desirable to present, the article needs all the benefit it can derive from a careful and obvious organization, and I have sacrificed fluency to a quite formal arrangement under headings and sub-headings.

which it is found expedient to ascribe to the atoms. These experiments are enormously important, for they invest the atom with a reality which nothing else could give it. To some they have given the hope that all the properties of the atom may one day be demonstrated unquestionably by direct evidence. There is little reason to expect that we shall see that day. The atom is no longer entirely a product of the scientific imagination; but neither is it entirely an object of experience. Most of its properties are invented, not discovered. Whether this invented and imagined entity is "real" is a difficult question. Perhaps it is best to evade such a question by asking the questioner what he means by "reality". As a matter of fact, it is not possible to discuss atomic theories thoroughly without raising and settling such formidable questions as, what is a theory? and, what is an explanation? and even, what is reality? perhaps eventually, what is truth? I do not aspire to answer these questions. But there are some common misconceptions about atoms which it is prudent to clear away at the beginning.

In the first place, one does not utter an atomic theory by saying that a substance is made up of small pieces, each exactly like a large piece of the substance in every respect except size. We should achieve nothing by saying that iron is made of black lustrous conductive magnetic atoms, or that glass is built of colorless transparent brittle insulating atoms, or that an apple consists of white soft sweet juicy atoms. The atoms must be endowed with fewer and simpler properties than the substance they are meant to compose, else they are futile. One must select some of the properties of the substance to be attributed to its atoms, and set the others aside to be explained by those.

Again, it is not obvious which properties should be selected for the atom; these depend largely on the fantasy of the atom-builder. However, certain qualities such as viscosity and plasticity, conductance for heat and conductance for electricity, opacity and transparency and lustre, warmth and flavor and fragrance, are not usually assigned to atoms. In general, the more a quality of a substance varies with its state, the less it is suited to be made an atomic quality. Ferromagnetism is a quality which one would assign almost instinctively to the iron atom; but it is possible to deprive iron altogether of this quality by a simple heat treatment, and hence it is not generally supposed to be a feature of the atom. But the rule is not an absolute one. The visible radiations from gaseous iron are supposed to be characteristic above all other things of the atom itself, yet they cease when the iron is condensed. It is supposed that in the condensed phases the atoms are so close together that they distort one another—a permissible idea if used with discretion, yet an atomic theory could easily become a meaningless form of words if this device were employed without limit. Of all the properties of matter, mass alone appears to be entirely exempt from change. For this reason all atom-models involve mass as an essential property of the atom; and this is the only respect in which they all agree.

Few and simple, therefore, must be the properties of the atom; yet we must not rush to the other extreme, and contrive atoms simplified into uselessness. The chemists know of eighty-eight different elements, sufficiently unlike to be distinguished; and we all know how great is the contrast between carbon and gold, hydrogen and lead, fluorine and helium. It is scarcely likely that such differences as these can be explained by atoms which are simply hard pellets differing only in size and shape and weight, like those of Lucretius and Newton, or by atoms which are abstract centres of force, like those of Boscovich. We are forced to invent atoms more complicated that these; and from this it is not far to say that we must imagine a structure for the atom; and from this scarcely farther to say that we must imagine an atom built of parts.

At this point we meet with a clamor from a number of excellent people, many of them otherwise quite innocent of Hellenic culture, who have it firmly fixed in their minds that *atom* is the Greek word for *indivisible*; whence they conclude that when the physicist speaks of subdividing his atoms, he is contradicting his own terms, he is violating the rules of his own game, and has forfeited his right to be heard.[2] The premise may be right, but the conclusion is absurd. If some of the properties of gold are explained by assuming it made of atoms with fewer properties, and later the explanation is improved and extended by assuming these atoms made of still smaller particles with still fewer properties, the second step is not less legitimate than the first. It may be contended, with some reason, that the name *atom* should be transferred at once to the smaller particles. At best this would be one of the changes which are desirable in principle but cause more trouble than they are worth. The contention is, however, weakened by the fact that some at least of the smaller particles of which we imagine gold atoms to be made are not imagined to be peculiar to gold, but are conceived as particles of a fundamental substance common to all elements. That the "atoms" of the many

[2] I should have put this passage even more strongly, but that Schuster tells that Kelvin himself inveighed on one occasion against the idea of subdividing atoms. He was answered by a young man who said, "There you see the disadvantages of knowing Greek." This seems as good an answer as any.

elements should be systems of "atoms" of one or a few fundamental materials is a thoroughly pleasing idea, although at present an unrealized ideal. It is unknown how far our descendants will find it expedient to dissect the atom; but it is certain that they will not be stopped by etymology.

Another fact about atom-models is that they are not always displaced by their successors; several may and do persist side by side, each adapted to a certain set of facts and observations. Every atom is designed in view of a very small fraction of the available knowledge about properties of matter; and this applies to the latest model as well as the earliest. The chemists of the nineteenth century were most impressed by the immutable weight of matter and by the laws of chemical combination; hence their atoms were merely weighted particles equipped with hooks to catch the hooks of other atoms. To the physicists of fifty years ago the physical properties of gases seemed the easiest phenomena to interpret, and they imagined atoms as rigid elastic spheres with radii of some 10^{-8} centimetre; by the masses and motions of such atoms they explained the pressure, elasticity, viscosity, diffusion and specific heats of gases. The physicists of the next generation attended chiefly to the emission, the refraction, the dispersion of vibratory radiations by luminous gases, and conceived the atom as a framework holding vibrators, like a belfry with a carillon of bells. This third model is inferior to the second in explaining the properties of gases, inferior to the first in explaining the laws of chemical combination; each of the three is superior in its own field to the atom-model to which this article is chiefly devoted, and which in its turn is primarily adapted to a field of its own. Still other atom-models have been devised, endowed with other properties, to account for other phenomena; and it is altogether probable that many more will be presented before the eventual one is attained, if it ever is. For instance, we may some day behold an atom-model devised to explain the conduction of electricity in solids, very competent in its field and quite unlike these others. In the eventual atom-model the essential qualities of all of these, and of many others, must be happily combined; it does not matter about the inessential ones.[3]

[3] Now and then an article appears in a physical or chemical journal, in which an oddly unconventional atom-model is proposed to interpret some such property of matter as the thermoelectric effects, or supra-conductivity, or valence, or some other with which the Rutherford-Bohr atom-model has not as yet been matched. It is easy for a physicist to ignore such articles, on the ground that any model departing from that of Rutherford and Bohr must be wrong. This is certainly a mistaken policy. Any partially competent atom-model deserves to be examined with care; its essential features must reappear in the eventual model. But, of course, the essential feature is not always the conspicuous one.

In awaiting that eventual atom-model, it is best to regard the atoms of the present day as mutable and transitory. Like railway time-tables, atom-models should be inscribed "subject to change without notice." Nothing is irrevocable in physics, except the record of past events; and we who have seen the undulatory theory of light assailed and shaken may well hesitate to put unquestioning faith in any atom-model. Even if there is no danger of change, it is a virtue to keep data and theories sharply separated in one's mind. In no field is this more difficult and important than in the field of this article, where the very language used to describe the data is saturated with the spirit of a particular conception of the atom, and it is customary to expound the theory before the facts. For these reasons I shall go to the opposite extreme, and treat the contemporary atomic theory with an exaggerated reserve which in many places will seem excessive to the reader and in some to the writer himself.

The favorite atom-model of the physicists of today is a structure of electrons, congregated about a positively-charged nucleus. The data which this atom is designed primarily to interpret were discovered before 1913, or else since 1913 by methods developed before that time. These discoveries are due largely to Rutherford, whose name the model often bears. The sections of this article which are labelled *B*, *C* and *D* are devoted to these data, and to the inferences from them. In 1913 a great change in the situation was wrought by a brilliant idea of Niels Bohr. Bohr did not discover new data; he taught a new way of interpreting old ones, he showed men how to read spectra. Through this interpretation of spectra, and through data which were discovered by men inspired with his idea, a previously-unknown property of matter was disclosed. This is expressed by saying that each atom possesses many distinct Stationary States. The largest section of this First Part of the article, the section *E*, is devoted to the knowledge of these Stationary States. Had these been discovered earlier, an atom-model might have been devised to explain them and them alone. Rutherford's atom-model was already in the field, and it was modified so that it might interpret the new knowledge. To these modifications, of which some are of a remarkable simplicity and beauty, the Second Part of this article will be devoted.

B. The Electron [4]

The *electron* is the atom of negative electricity. An individual electron can be captured upon a droplet of oil or mercury, or a minute

[4] This section is drastically curtailed, for the chief facts about the electron should by this time be common knowledge. Millikan's book "The Electron" (now in its second edition) may be consulted.

solid particle, and the amount of its charge determined. This amount is $4.774.10^{-10}$ in electrostatic units, according to Millikan. It is designated by the symbol *e*. When a magnetic field is applied to a stream of electrons all moving with the same speed, the electrons are deflected all to the same degree, which shows that they all have the same mass. This mass is practically equal to 9.10^{-28} in grammes, unless the electron is moving at a very uncommonly high speed, in which case it is appreciably greater.

These facts of experience are about all that is definitely known or needs to be known about the electron, in order to appreciate its role in modern atomic theory. There is no good way of determining its size, although the length of its mean free path in certain gases indicates, perhaps definitely proves, that it is much smaller than an atom. If the electron is a spherule of negative electricity uniformly dense, then its radius cannot be less than 2.10^{-13} cm, for if it were, the electromagnetic mass of the spherule would exceed the observed mass of the electron.[5] This size is much smaller than the one which it is expedient to attribute to the atom, happily for us, since otherwise it would be difficult to conceive of atoms containing electrons.

Since electrons can be coaxed or forced out of substances of every kind—elements and compounds, metals and non-metals, liquids and solids and gases—the atoms are supposed to contain one or more electrons apiece. This argument was formerly fortified by the fact that the light emitted from glowing gases is in many respects such as oscillating electrons would emit. This second argument is for the present under a cloud.[6]

[5] This is a short way of saying that, if the electron were a particle of smaller radius than 2.10^{-13} cm., more energy would have to be supplied to it to increase its speed than is actually required. For, in order to set an electrified particle into motion, energy must be supplied to build up the magnetic field which surrounds a moving electric charge; this energy U is additional to the kinetic energy $\frac{1}{2}mv^2$ required to set the mass m associated with the charge into motion with speed v, and it may be regarded as the kinetic energy associated with an extra "electromagnetic" mass $2U/v^2$ which the particle possesses by virtue of its charge. This quantity $2U/v^2$ can be calculated, for a given size and shape of the electron; if we make the electron too small, $2U/v^2$ comes out larger than its observed mass, which is a *reductio ad absurdum*. This illustrates the rather surprising fact that we are not permitted to imagine the electron as an infinitely small particle, a mere geometrical point loaded with an infinitely concentrated charge and mass. Speculations about its size and shape and the distribution of charge within it are not necessarily trivial; some may even be verifiable. We also meet with this dilemma: how does the electron, a piece of negative electricity of which each part should repel every other, keep from exploding?

[6] Perhaps I ought to mention that F. Ehrenhaft of Vienna has been ardently contending for about fifteen years that there is no such thing as an electron. He maintains that he can demonstrate negative charges much smaller in amount than

C. Positively-Charged Particles Accepted as Atoms[7]

Positively-charged particles are found in abundance in gases in which an electrical discharge is or has lately been maintained, and they may be produced under well-controlled circumstances by pouring a stream of electrons with properly-adjusted speeds into a gas, and in other ways. Only the ratio of the charge to the mass can be determined for these particles, not the charge individually nor the mass individually. But particles of apparently the same substance show distinct values of this ratio, which stand to one another as the numbers 1, 2, 3, . . . and the intermediate values do not occur. This supports the quite natural idea that these particles are atoms which have lost one or two or three or more of their electrons. If we make this supposition, we thereby assume values for the charges, and can calculate the masses of the particles from these and the observed values of the charge-mass ratio. The masses lie between 10^{-24} and 10^{-21} (in grammes) for particles occurring in the vapors of the various chemical elements, and they lie in the same order as the combining-weights of the chemical elements. This is powerful testimony that the particles indeed deserve the name of "atoms".

There is one sort of positive particle for which the charge can be measured directly. This is the alpha-particle, which cannot be produced at will but is supplied by Nature from radio-active substances. Counting the number of these particles emitted from a bit of radio-active substance in a given time, and measuring the total electrical charge lost by the substance in the same time, and dividing the latter figure by the former, Rutherford and Regener obtained the charge of the alpha-particle, which is twice the electron-charge (with reversed sign) within the limits of experimental error. This suggests that the alpha-particle is an atom of something or other, which has lost two electrons. As an evacuated tube into which alpha-particles are admitted is presently found to contain helium, the "something or other" is supposed to be helium. The mass of the alpha-particle can be determined directly from its charge and charge-mass ratio. It amounts to $6.60 \cdot 10^{-24}$, and this agrees with the mass inferred in the foregoing way for the positive particles found in helium.

$4.774 \cdot 10^{-10}$. Anyone interested in his case may find it presented in the April, 1925, number of the *Philosophical Magazine*. The question is for experimental physicists to discuss; but it is not likely that the edifice of modern physics is liable to be ruined by a flaw at its very foundation, such as this would be.

[7] The material of this section may be found much more extensively presented in my fourth article, in which I have also written about isotopes, a subject omitted here for the sake of brevity.

The alpha-particle is supposed, like the electron, to be much smaller than an atom; partly because it can go through a thin sheet of metal, chiefly because of evidence to be expounded in the next paragraph.

Collisions between alpha-particles and other particles of similar mass are occasionally observed; the mass of the struck particle can be deduced from the directions in which it and the alpha-particle fly off after the impact, assuming only that conservation of momentum and conservation of kinetic energy prevail during the impact. In this way it is possible to determine the masses of tiny particles (presumably atoms) of hydrogen, helium, oxygen and nitrogen (perhaps eventually of other elements) in terms of the mass of the alpha-particle, which is determined from its charge-mass ratio and its charge, which are determined directly. If all the properties of the elements could be explained by atoms possessing no features except charge and mass, all the foundations of science might be laid down already.

The alpha-particle is one of the most valuable and powerful instruments in the physicist's equipment. It is a sort of hyper-microscope, penetrating and revealing the arrangements of systems so minute that microscopic objects are universes compared with them. Rutherford's development of the technique of using the alpha-particle is to be ranked among his greatest works.

Positively-charged particles with masses as low as that of the electron have never been observed; the least massive of the known positively-charged particles has 1,840 times the mass of the electron.

D. The Nuclear Atom-Model

Since we have met with positively-charged particles which are accepted as atoms deprived of one or more of their electrons, and since these incomplete atoms are much greater in mass than the electrons, it is natural to suppose that the completed atom consists of a positively-charged particle or *nucleus* in which almost its entire mass is concentrated, and one or more electrons which compensate the charge of the positive particle but add little to the mass of the atom. If we further suppose that the dimensions of the electrons and of the positively-charged particle are small in comparison with the distance between them, we invent the *nuclear atom-model*.[8]

The direct evidence for the nuclear atom-model consists of a very

[8] Commonly known as the Rutherford atom-model, after the physicist who invented it and discovered most of the evidence for it; occasionally as Nagaoka's, after another physicist who suggested it; occasionally as the Saturnian model, as some have supposed that the electrons lie in flat rings around the nucleus like the rings of Saturn around that planet.

small but a beautiful and convincing series of experiments, of which the first and the most were performed by Sir Ernest Rutherford and his pupils.[9] These experiments are designed to show that the orbit of a minute charged particle (usually an alpha-particle), flying through a thin film of metal, is in certain cases very like the hyperbolic orbit of a comet around the sun. Such an orbit is the path of a particle moving near to an immobile particle, for instance a light particle moving close to a much more massive one, which attracts it or repels it by a force varying inversely as the square of their distance apart. If these experiments show what they are designed to show, then they indicate that the atom includes a particle much more massive than an electron, bearing an electric charge, and sufficiently isolated from the other charges in the atom (such as the electrons) so that its field of force in a measurable space around it is not disturbed by theirs. We cannot, however, trace the entire path of an individual flying charged particle as it swings around through an atom, and are forced to make up for this deficiency by a statistical study of the visible portions of the paths of a great multitude of charge particles.

Let us consider exactly what these experiments show; for whatever they do prove is the most securely proved of all the beliefs about atoms. In the first place, they show that there is a nucleus; and a vacant space surrounding it, in which an inverse-square force centred upon the nucleus prevails; and they indicate the dimensions of this vacant space. This commences within 10^{-12} cm. of the nucleus, which is another way of saying that the diameter of the nucleus is less than 10^{-12} cm.; and it extends beyond a distance given (to take instances) as 14.10^{-12} cm. for platinum and 10^{-9} cm. for argon, which is another way of saying that nearly all of the negative charge of the atom lies still farther out from the nucleus. If the negative charge is indeed subdivided into electrons, then the atom is formed like a hollow cloud of electrons, with a massive positively-charged nucleus at the centre of the interior hollow.

The diameter of this cloud of electrons is not furnished by the experiments on alpha-particle deflections; but considering that the distance between adjacent atoms locked into a crystal lattice is generally a small multiple of 10^{-8} cm., it cannot be much greater than 10^{-8} cm. unless we are prepared to admit interpenetration or violent distortion of atoms; nor does it seem likely that the diameter is very much smaller than this amount. I have already mentioned that some of the properties of gases are adequately explained by assuming

[9] For the mathematical theory of these experiments, the second article of this series may be consulted.

that the atoms are elastic rigid spheres with a diameter of about 10^{-8} cm. Unlike as an elastic rigid sphere and a cloud of electrons seem, this agreement between so differently made estimates is probably no mere coincidence. It will be noticed that all of the figures about sizes at which we have arrived in such various ways (diameters for the electron and the nucleus, for the vacant space inside the electron-cloud, for the entire atom) are quite compatible with one another. If the value derived for the diameter of the interior hollow had been ten times the spacing of atoms in a crystal, or a tenth the diameter of a spherule of electricity with the same electromagnetic mass as an electron, we should indeed be in trouble.

In the second place, these studies of the deflections of alpha-particles yield numerical values for the nuclear charge: $(77.4 \pm 1)e$ for platinum, $(46.3 \pm 0.7)e$ for silver, $(29.3 \pm 0.7)e$ for copper, $19e$ for argon, $6.5e$ for "air" (a sort of statistical average of the values for oxygen and nitrogen).[10] To these must be added the value $+2e$ for the nuclear charge of helium; for we have already seen the evidence that the alpha-particle is what is left of a helium atom when two electrons are removed, and these last-cited experiments show that it is itself a nucleus, hence a helium nucleus. This nuclear charge must be balanced by negative charges within the atom; if this balancing negative charge is subdivided into electrons, then the numerical factors of e occurring in these numerical values are equal respectively to the number of electrons belonging to each atom. We thus have fairly accurate estimates of the number of constituent electrons within each of four or five atoms.

These estimates agree, within their experimental uncertainties, with the famous and splendid idea of van den Broek and Moseley: that the number of electrons in each atom, and the nuclear charge measured as a multiple of the electron-charge, "is the same as the number of the place occupied by the element in the periodic table". This idea is also supported by rough measurements of alpha-particle deflections by a few other atoms, and by the extent to which different atoms scatter X-rays; but the most important of the additional evidence will find its appropriate place in the second section of this article.

These conclusions are almost all that can be deduced from the data. The arrangement of electrons within the electron-cloud is almost

[10] References for these data are given in the fourth article of this series. The data obtained by E. S. Bieler (Proc. Roy. Soc., 105A, pp. 434–450, 1924) show incidentally, if I do not misread his article, that the nuclear charges of Mg and Al have the desired values $12e$ and $13e$, respectively, within a few per cent. Rutherford's studies of encounters between alpha-particles and hydrogen atoms prove a nuclear charge of e for the latter.

entirely concealed. It is not altogether inaccessible; for the deflections suffered by alpha-particles and electrons flying through atoms are influenced by the electrons of the atom, not by the nucleus exclusively; and from the degree in which the observed deflections differ from what the nucleus alone would compel, it is possible to draw some conclusions about the way in which the electrons are arranged. The mathematical difficulties, as the reader will readily admit, are tremendous; the problem of determining the path of a flying electron through a cloud of electrons, probably themselves in motion, is enough to make the best of mathematicians despair; yet some progress in this direction has already been achieved, as I narrated in the second article of this series. Again, the scattering of X-rays by atoms should depend on the manner in which their electrons are arranged; and some measurements and some deductions have already been made, although the researches have been in abeyance for some years, probably because the newest discoveries about X-ray scattering make it extremely doubtful what the mechanism of the effect really is.

The study of deflections of alpha-particles by atoms has thus brought precious guidance to the atom-builder, and imposed severe limitations upon him, yet only partial ones. He is constrained to erect his atom according to certain fundamental rules, and yet has an extremely free hand in arranging the details. He is practically compelled to build the atom of an element which occupies the Nth place in the periodic system, out of N electrons and a much more massive nucleus with a positive charge Ne. The data which I have cited do not absolutely enforce these numerical values; but there is no other model which they permit which could possibly rival this one in respect of convincing simplicity. He may not make the electrons go more than a few times 10^{-8} cm. from the nucleus; he is constrained to leave a small vacant space around the nucleus, and within this space he may not tamper with the inverse-square law of force (a restriction which has eliminated several favored atom-models of the decade before 1910).[11] Having conformed to these restrictions he

[11] Except that he may and must alter the inverse-square law of force to just the extent that further and more delicate experiments of this type require. Thus Bieler (*l.c. supra*) concludes, from a study of deflections of alpha-particles passing close to the nuclei of aluminium atoms, that within about 10^{-12} cm. of the aluminium nucleus the inverse-square repulsion which it exerts upon an alpha-particle is supplemented by an attractive force—perhaps an inverse-fourth-power attraction, just balancing the repulsion at a distance of $3.44 \cdot 10^{-13}$ cm. from the centre of the nucleus. Rutherford earlier found anomalies in the encounters between hydrogen nuclei and alpha-particles, which suggested to Darwin that the latter might be considered as a disc-shaped hard particle, or an oblate spheroid of semi-axes 4.10^{-13} and 8.10^{-13} cm.; this would repel hydrogen nuclei according to the inverse-square law so long as it did not actually strike them.

may do very nearly as he pleases with the electrons and the region they occupy. No data can be invoked to support him nor to confute.

Having expounded the merits of the nuclear atom, I will proceed to undo my work in part by pointing out its great and grave defect. No less a defect than this, that it is impossible. It cannot exist. Even if it were brought into existence miraculously at an instant, it could not survive, for it carries the seeds of its own dissolution within itself. For if at that initial instant all of the electrons were at rest relatively to the nucleus, they would immediately start towards it, fall into it, and expire. Of course, this consequence is so obvious that the notion of stationary electrons would not even occur to anyone having a bowing acquaintance with mechanics. Such a person would immediately assume that the electrons were in motion around the nucleus as the planets are around the sun; he would convert the nuclear atom-model into what I might call a *sun-and-planets atom-model*, the nucleus playing the role of the sun, the electrons those of the planets. Such an idea is alluring in the extreme; it implies that Nature acts similarly in great things and in small, copying the solar system within the atom; and this is most acceptable, partly because of its philosophical beauty and partly because it enables us to use the intellectual methods and habits acquired in the study of astronomy, relieving us of the labor of acquiring new ones. Unfortunately it is as untenable as the idea that the electrons stand still. For owing to the radiation of energy which continually goes on from accelerated electrified particles, an electron cannot describe a circle or an ellipse about a nucleus, as a planet may about the sun; it can only describe a narrowing spiral, ending in a collision between electron and nucleus. The nuclear atom is not stable nor enduring; and the dilemma is complete.

The only recourse is to make some entirely new and unprecedented assumption; for instance, that the electrons, in spite of everything, can stand still in certain positions without falling into the nucleus; or that they, in spite of everything, can revolve interminably in certain closed orbits without spiralling into the nucleus. Such a modification of the nuclear atom is, of course, essentially a denial of it. An atom composed of masses and electrostatic charges, plus certain restrictive rules or arbitrary assertions, is no longer simply an atom composed of masses and electrostatic charges. Instead of giving to our ultimate particles a few properties selected from among the ones which matter *en masse* displays to our senses or our instruments, we have to invent some new ones for them. This seems regrettable, but only because our expectations were too high.

Another circumstance leads us to another dilemma. Suppose that we could circumvent that difficulty about the revolving electron, which radiates part of its energy at each revolution and slides down a spiral path into the nucleus; suppose that we could find justification for saying that no radiation occurs, that the electron like a planet may revolve forever in an ellipse. If two atoms collided, as in a gas they must very frequently do, would not the electrons all be disarranged, disorganized, flung over from one orbit into another? This we should certainly expect; yet if it happens, no two atoms in a gas can be exactly alike, nor can any atom retain its character for more than a fraction of a second. If this is so, then the various sharply-defined properties of a gas must, each and every one of them, be statistical properties—not themselves properties of individual atoms, but the results of other properties of individual atoms, held in different amounts by different atoms and all averaged together. In some cases this is unobjectionable; the pressure and the temperature of a gas are sharply definite properties, resulting from the mass and the motion of the atoms, and the latter of these properties is not necessarily the same for any two atoms at the same moment nor for any atom at different moments. But one would be reluctant to treat the spectrum of a gas as such a property; according to all the traditions of physics this is one of the properties of the individual atoms. But the spectrum is very constant, sharp, immutably defined; we must therefore assume either that it depends only on the number of electrons in the atom and not upon their motion nor position, an idea which would be difficult to carry through; or that the electrons are ineluctably constrained to certain orbits or certain positions, so that the atom retains its personality and its character.

We have now made the acquaintance of two ideas which will be exceedingly prominent in the second division of this article. The nuclear atom-model is of itself unstable; therefore stability must be enforced upon it by outright assumption, it must be made stable by fiat. But this stability may not be extended to all conceivable arrangements or configurations of the model; it must be reserved for one or a few, that the atom may possess a fixed character and a personality.

We now arrive at the phenomena by means of which these vaguely-expressed ideas are to be sharpened and hardened into definite doctrines.

E. The Stationary States

E 1. *The Direct Evidence for the Stationary States*

Imagine a tube filled with gaseous helium, and containing a hot filament from which electrons emerge. By means of an accelerating potential applied between the filament and a fine-meshed gauze close in front of it, the electrons are speeded up, and pass through the gas with an energy which is accurately controlled by the accelerating-potential. A third electrode is maintained at a potential only slightly higher than that of the filament. To reach this electrode, the electrons must sacrifice nearly all of the energy which they acquired in coming up to the gauze. If they lose little or no energy in their progress through the gas, they can win their way to the third electrode, like water rising again to the level of its source. If, however, they lose a notable amount of energy to the atoms with which they collide, they cannot reach the third electrode, as water which has turned a mill-wheel cannot climb again to the level whence it fell.

By measuring the current into the third electrode in the helium-filled tube, it is found that if the electrons have an amount of energy lower than 19.75 equivalent volts, they lose scarcely any of it in their progress through the gas; but if the energy of an electron is just equal to 19.75 equivalent volts, it may and frequently does lose its energy altogether; and if the energy of an electron surpasses 19.75, it may and frequently does surrender just 19.75 equivalent volts to the gas, retaining the residuum itself. Imagining that the electron collides with atoms of helium on its way across the gas, we conclude that the helium atom can receive exactly 19.75 of these units of energy, no lesser quantity and (within certain limits) no greater. From similar experiments it appears that the mercury atom can receive 4.66 equivalent volts of energy, no smaller amount and (within certain limits) no larger. It appears that the sodium atom can receive 2.1 equivalent volts, no less and (within certain limits) no more—and the list can be extended to some thirty elements.

Another way of saying the same thing is this: the helium atom may exist (at least transiently) in its normal state, *or also* in a second state in which its energy is greater by 19.75 equivalent volts than in its normal state,—but not, so far as we can find evidence, in any state with any intermediate value of energy. Let us call this second state an "excited state." The mercury atom then has, in addition to its normal state of undefined energy, an excited state of energy greater by 4.66 equivalent volts. The sodium atom has, in addition to its normal state, an excited state of energy greater by 2.1 equivalent volts and so with a number of others. I give these and a few other values in the following table:

TABLE I

	He	Ne	Na	Cs	Mg	Hg
Energy-value of the						
Normal state	0	0	0	0	0	0
First excited state	19.75	16.65	2.1	1.45	2.7	4.66
Other excited states.	20.55	18.45			4.4	4.86
						5.43
						6.7
Ionized atom	24.5	21.5	5.12	3.9	7.6	10.4

It will be noticed that values are given for several excited states in the same column; these rest upon evidence of the same sort as does the first excited state, so that in general the atom must be considered to possess not one only, but several possible states in addition to its normal state.

It will be noticed also that values are given for the "ionized atom." These are the amounts of energy just sufficient (when applied by means of an impinging electron) to detach an electron from the atom. When electrons with so much energy or more are poured into the gas in question, positively-charged particles, such as I previously mentioned and characterized as the residues of atoms deprived of an electron apiece, appear in it. It is not absurd to call this an "excited state." If it takes just 24.5 equivalent volts of energy to detach an electron from a helium atom, then the system formed of an ionized helium atom and a free electron has a potential energy of 24.5 equivalent volts. Any experiment, therefore, in which the energy required to detach an electron from an atom is measured—any experiment for determining the *ionizing-potential*, as this energy when expressed in equivalent volts is called—is essentially an experiment for locating one of the excited states of the atom.

In this sense the energy-values of the last line in Table I are to be taken. I introduce them here for two reasons. In the first place, the fact that this energy-value is greater than any of the others in the same column suggests this interpretation for the excited states: that they correspond each to a certain *partial* lifting-out of an electron, to a certain stage of *incomplete separation*, while the energy-value of the ionized atom corresponds to the *total* lifting-out or to the *complete separation*. This idea is fortified by the fact that a helium atom may be ionized by two consecutive blows from electrons each with

20 equivalent volts of energy, if the blows fall closely enough together—as if the energy spent in raising the atom to its first excited state were paid into account, and could be used toward detaching the electron when the deficiency is supplied. This fact is exceedingly important for the theory, and I mention it here as a passing anticipation. In the second place it is desirable—for a reason which will presently appear—to measure the energy-values of the normal and of the excited states not from the energy of the normal state, as I have done in Table I, but from the energy of the ionized atom as zero-value. This is done in Table II.

TABLE II

	He	Ne	Na	Cs	Mg	Hg
Energy-value of the Ionized atom . . .	0	0	0	0	0	0
Non-ionized atom					−3.2	−3.7
Excited states. . . .	−3.95	−3.0				−4.97
						−5.54
First excited state.	−4.75	−4.85	−3.0	−2.15	−4.9	−5.74
Normal state. .	−24.5	−21.5	−5.1	−3.9	−7.6	−10.4

With this convention, all the energy-values for the non-ionized atom become negative—a source of confusion, but not of nearly so much confusion as the previous convention would eventually entail. It is well to remember tenaciously that, in at least nine cases out of ten in the literature, the energy-values of the normal state and the excited states are referred to the energy of the ionized atom as zero, and that they all should always bear the minus sign, though generally it is left off.

For the excited states and for the normal state, I will employ the common general name of *Stationary States;* occasionally, for the sake of variety, the alternative name *levels*. Another common word is *term*, the origin of which will appear in the next section.[12]

As the reader will be forced to make himself familiar with schematic representations of the Stationary States, he may as well begin at once with a simple one. Fig. 1 is a diagram showing the stationary states listed for helium in the foregoing tables. The levels are represented by horizontal lines, separated by distances proportional to the

[12] Anyone who reads the physical literature of today soon becomes familiar with the phrase "the electron is in the . . . orbit" used instead of "the atom is in the . . . state." This phrase expresses theory rather than facts of observation, and does not always express theory adequately; I have avoided it in this article.

differences between their energy-values (usually, however, these distances are distorted for convenience). The energy-values, expressed in equivalent volts, are affixed to the lines; on the left, they are measured from the normal state of the neutral atom as zero of

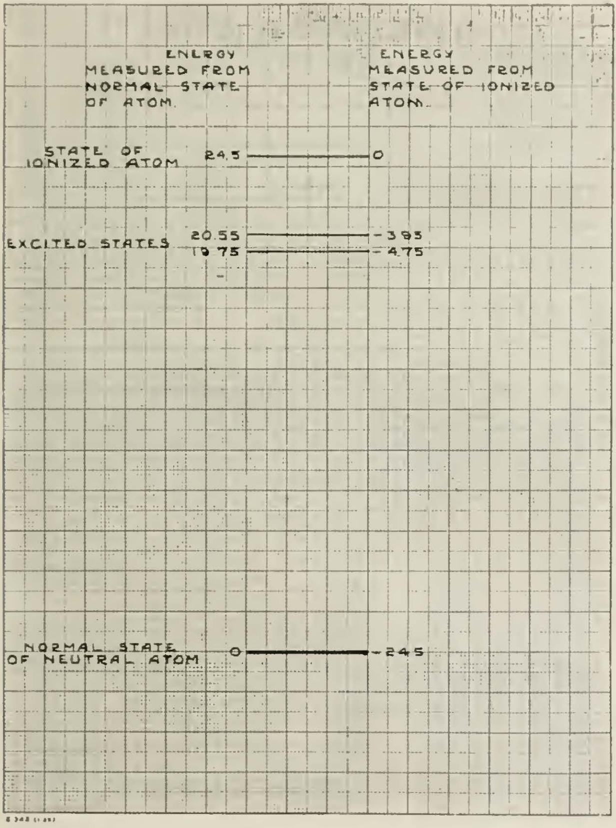


Fig. 1—Diagram of the stationary states of helium, determined by the method of electron-impacts

energy; on the right, they are measured from the state of the ionized atom (which is the more common practice)[13].

E 2. Bohr's Interpretation of Spectra

In 1912, the evidence to which the foregoing section is devoted was still entirely undiscovered, the Stationary States were unknown. That evidence was sought and found because Niels Bohr had divined the Stationary States in developing a new and brilliant interpretation of spectra. Until then, all physicists had wished to interpret the frequencies forming the spectrum of an atom as the natural resonance-frequencies of an elastic system. Bohr supplanted this idea with an idea of his own, one of the most novel, fecund and potent in all the long evolution of physics. Several of the ideas incorporated in the contemporary atom-model are due to Bohr; among them all this is the primary and fundamental one, and certainly the most secure.

Consider the spectrum of hydrogen. In the visible region, this spectrum consists of a "line-series"—that is to say, a procession of lines converging upon a limit, falling at intervals ever narrower and narrower, these intervals so smoothly diminishing that they bear witness to a common character and a mutual origin of all the lines.

Fig. 2—Balmer series of lines in the hydrogen spectrum. (R. H. Curtiss, from Foote & Mohler, "Origin of Spectra")

This line-series is shown in Fig. 2. Not only to the eye is it of a wonderful regularity; the frequencies of its consecutive lines are bound together in a simple numerical law. They are equal successively to

$$\nu_{lim}-R/3^2,\ \nu_{lim}-R/4^2,\ \nu_{lim}-R/5^2,\ \nu_{lim}-R/6^2, \text{ etc.} \qquad (1)$$

[13] This method of locating stationary states by observing transfers of energy from electrons at atoms is called the *method of inelastic impacts*; for the impacts of electrons against atoms are elastic (by definition) so long as there is no transfer of energy into the internal economy of the atom, and are inelastic when such transfers occur. When an atom returns into its normal state from an excited state, it usually emits radiation; hence a method for detecting the first commencement of radiation is usually (perhaps not always) equivalent to a method for detecting the first commencement of inelastic impacts. As it is generally easier to set up apparatus for detecting radiation than to seek evidence for elastic impacts, direct observations upon these last are not so abundant as they should be. Nobody really knows how many stationary states might be discovered by the method of inelastic impacts, although Franck and Einsporn detected over a dozen for mercury (of which those given in Table II are some). In fact they detected more than could conveniently be ascribed to mercury atoms, so that it was necessary to attribute some of them to molecules.

in which

$$\nu_{lim} = \text{frequency of the limit of the series} = R/4$$

R standing for a certain constant. There is another series of lines in the ultraviolet part of the same spectrum, whereof the frequencies are equal consecutively to

$$\nu = \nu_{lim} - R/4,\ \nu_{lim} - R/9,\ \nu_{lim} - R/16, \text{ etc.} \tag{2}$$

in which

$$\nu_{lim} = R,$$

R having the same value as before. The utter simplicity of the terms to be subtracted from ν_{lim} in each of these cases, not to speak of the related form of the expressions for ν_{lim}, suggests like simple laws in other fields of physics that in this formulation of the facts something highly important has been partially unveiled. There are certain other series in the spectrum of hydrogen, and inspecting them all one is led to the rule that *every frequency emitted by the hydrogen atom can be calculated by inserting different pairs of integers in the places of* m *and* n *in the formula*

$$\nu = R\left(\frac{1}{m^2} - \frac{1}{n^2}\right). \tag{3}$$

The case of the ionized-helium [14] atom is quite as simple. Every frequency emitted by this atom can be calculated by assigning different pairs of integer values to the constants m and n in the formula

$$\nu = 4R\left(\frac{1}{m^2} - \frac{1}{n^2}\right). \tag{4}$$

Line-series have been found in the spectra of many other elements. Some of them are as strikingly outstanding as the line-series in the spectrum of hydrogen, and converge upon limits scarcely less easy to locate; for instance, the "principal" series of the spectrum of sodium (Fig. 3). Most are by no means so obvious; often they are involved in the midst of a luxuriant jungle of unrelated or otherwise-related lines. Most spectra conceal their structures from the unpractised eye, as a tone-poem of Strauss its themes or an opera of the Ring its *Leitmotiv* from the inexperienced ear. Long training and a skilled judgment are required in the deciphering of spectra, except in the few untypically simple cases; and usually the arrangement of lines into series which the spectroscopist presents must be

[14] The reader may take this, for the time being, simply as the name of a particular element.

accepted by the theorist without question and without suggestion, for he is not competent to analyze the data for himself.

Having grouped a certain number of lines into a series, having guessed as well as possible the convergence-frequency ν_{lim} of this series, the spectroscopist has still the task of finding the numerical

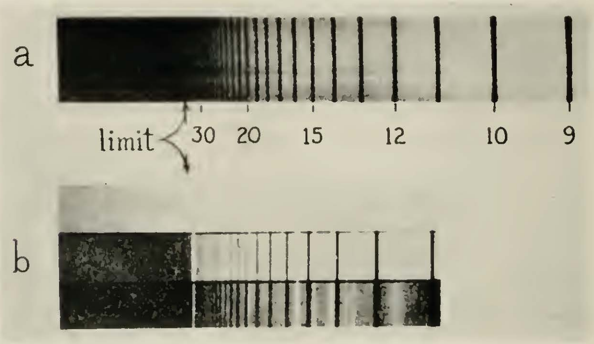


Fig. 3—Principal series of sodium (two photographs). (G. R. Harrison, *Physical Review*)

law to which the consecutive frequencies conform. As a matter of course, all the frequencies can be expressed by a formula generalized from (1) and (2):

$$\nu_i = \nu_{lim} - f(i) \tag{5}$$

in which i is the order-number distinguishing each line, and $f(i)$ is a different quantity for each of the lines, which approaches zero as we pass along the series to the limit. This means nothing by itself; the question is, does the function $f(i)$ have a simplicity comparable with the simplicity of the subtrahenda in (1) and (2) which suggested that they are the symbols of something deeply important? In general, the function $f(i)$ is not so simple as the function which occurs in the series of the spectra of hydrogen and ionized helium. In many cases, however, it is almost as simple, in others a little more complicated, in others a little more complicated yet, and so forth; so that the eventual result is this, that the formula (3) appears to be the proper way of describing the lines of series spectra, even in cases where the series is so irregular and the form of the function $f(i)$ so intricate that if it were the only series in existence, no one would attach any particular importance to it.[15]

To the physicists of a generation ago, who regarded the spectrum frequencies as natural vibration-frequencies of the atom, and tried hard to invent a mechanical model of which the vibration-frequencies should conform to the formula (3) or the more general formula (5), the character of these formulae was an insurmountable obstacle. Elsewhere [16] I have given a brief account of the vain attempts to contrive such a model. Bohr abandoned this procedure altogether; and taking equation (3), he multiplied both sides of it by Planck's constant h ($=6.56 \cdot 10^{-27}$).

$$h\nu = hR\left(\frac{1}{m^2} - \frac{1}{n^2}\right). \tag{6}$$

The significance of this act depends on the meaning of h. Planck had found it expedient, in developing an adequate theory of radiation, to assume that solid hot bodies are populated with multitudes of

Fig. 4—Principal series of helium (singlet system). (T. Lyman, *Astrophysical Journal*)

oscillating electrons of all the various frequencies, possessing a very curious and inexplicable property; this being, that an oscillator vibrating with frequency ν can emit radiant energy of that same frequency ν only in units or *quanta* of amount $h\nu$. Einstein had found it expedient, in describing the photoelectric effect and other phenomena, to assume that radiant energy of the frequency ν goes about in units or quanta of the amount $h\nu$, emitted integrally, absorbed integrally, travelling integrally. Suppose then that we assume that the quantity $h\nu$, standing on the left-hand side of the equation (6), represents the amount of radiant energy emitted by the hydrogen

[15] As a matter of fact, the series-limit is not generally so obvious to the eye that it can be located at once; it is determined after and by means of a careful choice of the most suitable form for the function $f(i)$. This is one of the difficulties of the spectroscopist's task.

[16] In the seventh article of this series (footnote 9).

atom in the process of pouring out radiation of the frequency ν. The right-hand side is the difference between two terms. One term is the energy of the hydrogen atom before it emits the radiation of frequency ν; the other is the energy of the atom after the emission is concluded. *The radiation of frequency ν is emitted by reason of a transition between two stationary states of the hydrogen atom; the energies of these states are equal to the terms whereof the frequency ν is the difference, each term multiplied by h.* The terms of the spectrum formulae are the energy-values of the stationary states of the atom, when translated into the same units by multiplying them by h. When translated into proper units, *the terms are energies, and the energies are terms.* This is Bohr's great and memorable idea.

Once this idea is accepted, the known stationary states of the atom increase enormously in number. The paltry one, two, or half-dozen, which are all that have been detected by observing the energy-losses of rebounding electrons, are multiplied into hundreds and thousands. The accuracy with which each energy-value is known is augmented tenfold or a hundredfold, sometimes far more; for spectroscopic measurements are among the most accurate in physics, although the necessity of extrapolating the observed frequencies to arrive at the series-limit neutralizes some of their precision.

One point must be kept clearly and always in mind, at the peril of infinite confusion. *The energy-values which the spectrum terms supply are* not the energy-values of the stationary states measured from the normal state, as might seem natural; they are *the energy-values measured from the state of the ionized atom.* These being negative, it is the negative term-value which is significant. Equation (6) must therefore be rewritten in this fashion:

$$h\nu = Rh\left(-\frac{1}{n^2}\right) - Rh\left(-\frac{1}{m^2}\right). \tag{7}$$

The energies of the successive stationary states of the hydrogen atom are $-Rh$, $-Rh/4$, $-Rh/9$, $-Rh/16$, and so forth, relatively to the energy of the ionized atom as zero. They are not Rh, $Rh/4$, $Rh/9$, and so forth, relatively to the normal state of the atom as zero. Anyone who entertains this last idea is doomed to trouble.

The stationary states of the hydrogen atom are shown in Fig. 5, which is constructed like Fig. 1, with the energy-values of the various levels measured downwards from the state of the ionized atom, and affixed on the right. The distances from the various levels to the zero-line are proportional to these energy-values (this feature will henceforth be found too inconvenient to maintain).

The energy-value of a stationary state, when obtained by analyzing a spectrum, is generally given not in equivalent volts, but in a unit called the "wave-number." This unit is $1/hc$ times as great as an erg, and $300hc/e$ (about .0001237) times as great as an equivalent volt. When the energy-values of two stationary states are expressed in

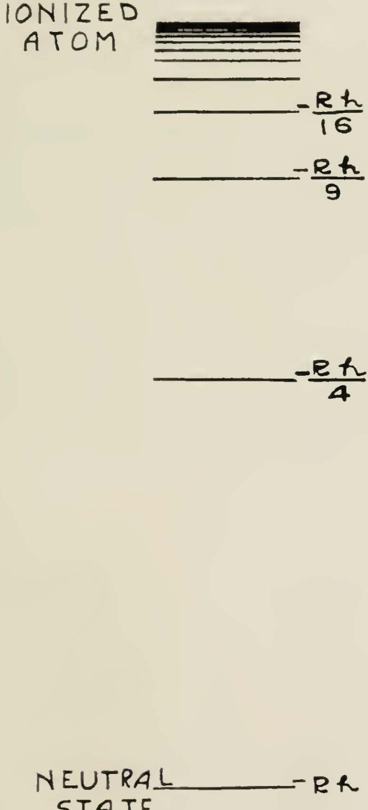


Fig. 5. Diagram of the stationary states of hydrogen, deduced from its spectrum

this unit, the difference between them is equal to $1/c$ times the frequency of the line which corresponds to the transition from one to the other.

A spectrum-line corresponding to a transition between two stationary states is symbolized, on a diagram of stationary states, by an arrow connecting the dashes (or whatever marks are used) which symbolize the two levels. This is illustrated in Fig. 6.

I pause at this point to remark that each of what I have been calling the "stationary states" is in fact usually a group of two or more stationary states, often but not always exceedingly close together; just as many stars in the sky are in fact groups of stars too close together for any but an excellent telescope to discriminate. This will be discussed at length in a later section; at present it is expedient to regard each of these groups as one stationary state.

The experimental test of Bohr's method for identifying stationary states consists in comparing the stationary states inferred from the spectrum, according to Bohr's procedure, with the stationary states derived directly by the study of electron-impacts. The agreement is perfect wherever the experiments by the latter method can be carried out. By a curious fatality, this is impracticable for hydrogen and ionized helium, as neither sort of atom occurs in gas quiescent enough for experiments on energy-transfers from electrons to atoms.

For about fifteen other elements, the comparison has been made for two or more of the Stationary States. Every energy-value given in Table II was obtained by the method of electron impacts, and confirmed by analyzing the spectrum of the element.

E 3. The Classification of Stationary States by Utilizing "Rules of Selection"

I have said that every line in a spectrum, at least of those arranged in series, may be represented by an arrow connecting two stationary states. If arrows are drawn from every one of the stationary states to every other, will every arrow correspond to a line actually observed in the spectrum? Every line has an arrow; does every arrow have a line? By no means; the answer is definitely and strongly negative. If the wave lengths deduced from all the possible arrows are sought in the spectrum, most of them are found unoccupied by lines. The great majority of the apparently possible transitions either do not occur at all, or if they do occur, the energy which is liberated is disposed of in some other way than by radiation. There is reason for believing that the atom may embrace this last alternative if it col-

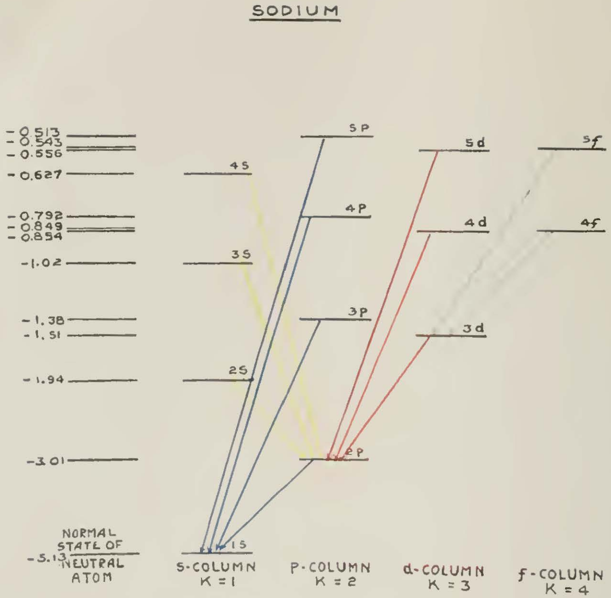


Fig. 6 Diagram of the stationary states of sodium, sorted out into columns by applying the selection-principle. Arrows represent various lines (blue for principal, yellow for sharp, red for diffuse and green for Bergmann series)

lides with another atom or with an electron. Otherwise, it seems that if the atom cannot radiate the energy liberated in a transition, the transition itself cannot happen at all. If, therefore, the line corresponding to an arrow is missing, the transition corresponding to the arrow must be inhibited by some agency as yet unknown. Many transitions must be inhibited, for many lines are missing.

These missing lines are precious to the student of spectra and to the architect of atom-models. Whatever explanation is devised for the stationary states must include a reason for the occurrence of some transitions and the non-occurrence of some others. This is good rather than bad fortune, since if such a reason is demanded, it may be found in one and not in another of two competing theories which otherwise would stand on an equal footing; the missing lines may even suggest a theory. At all events they suggest a system of classification; and, while the hardened theorizer may regard a system of classification as merely the forerunner of a theory, a theory is itself often nothing more than a classification stated in the language of an artificial analogy. It is, in fact, possible to arrange the stationary states, not in a single column as in Figs. 1 or 5, but in several as in Fig. 6; this arrangement being so contrived, that any transition can be identified in a moment as belonging among those which occur, or among those which are missing, whichever its case may be.

The mere fact that such an arrangement can be contrived shows that the missing lines are not distributed at random, but subject to some sort of a rule. Such rules are known as *principles of selection.* The missing lines are commonly called *verboten* lines by the German physicists, possibly because that was the most conspicuous word in the official German language before the war. It is not a happily-chosen word, neither are the English equivalents "forbidden" and "prohibited"; since while we know that the lines are missing, we do not definitely know what circumstance is responsible; and, whatever that circumstance may be, it is highly unconventional for a physicist to say that it "forbids" the lines. The same objection applies with extra force to the phrase "forbidden by the selection-principle". It is much better to accept the fact that certain lines are missing as a fact of experience, and the selection-principles as rules of experience whereby the facts are codified.

E 4. The Families of Stationary States (for Other Atoms than Hydrogen)

There is a far-reaching contrast between the spectra of all atoms but hydrogen and ionized helium, on the one hand, and the spectra of these two atoms on the other. The selection-principles at first accentuate this contrast, and later to a certain extent aid to explain it away. I commence with the atoms other than hydrogen, and take sodium as the specific instance.

A few of the stationary states of the sodium atom are exhibited in a single column on the left of Fig. 6. The energy-value of each level, measured from the energy of the ionized atom as zero, is affixed at the left; but the practice of drawing the levels at distances proportional to their energy-values has had to be discarded for the sake of lucidity. In this case, the distances are proportioned to the differences between the logarithms of the energy-values. Drawing arrows from each of the levels to every other, and ascertaining which of them correspond to actual and which to missing lines, we find that the missing lines are such that the stationary states can be sorted out into several families, to be arranged in parallel columns as on the right of Fig. 6. There are at least seven of these, but it is of no advantage to us to consider more than the first four. The feature of this arrangement is, that *transitions between stationary states in adjacent columns correspond to actual lines*; *but the lines corresponding to all other transitions are missing.*

This is a principle of selection. It may be phrased in an equivalent but pregnant way, in this manner. Let me attach to the several columns the numerals 1, 2, 3, 4 . . . , as they are indicated at the bases; and let me use k as the general symbol for each and all of these numerals. Then this particular selection-principle may be phrased thus:

> *The only transitions which correspond to actual spectrum lines are those in which* k *changes by unity*; $\Delta k = \pm 1$.

The numeral k bears the ponderous name of *azimuthal quantum-number*. This is a name derived from theory and not from experience, as will be made clear in due time. The principle of selection which has just been stated is the selection-principle for the azimuthal quantum-number.

Exceptions to this rule occur; the *verboten* lines, like other *verboten* things, occasionally evade the prohibition. This happens particularly when the atoms are subjected to intense electric fields, or to violent spasmodic electrical discharges in which strong transient fields are produced; in these circumstances great numbers of the missing lines leap suddenly into sight. In Fig. 11 some of these lines appear elicited by a strong electric field. Some lines corresponding to changes of k by two units or by none, which by the foregoing rule should be absent, do actually occur even when there is no obvious reason whatever for thinking that the atoms are subject to unusual stresses.[17] The exceptions, however, are not numerous enough to jeopardize the rule.

Two other features of the columns should be pointed out; first, that the successive levels in each column are not scattered at random, but form a converging series approaching the top of the column as limit (their energy-values form a sequence converging to zero); and second, that there is nothing arbitrary about the order of the columns, since the column at the extreme left admits of transitions to only one other column and therefore is unmistakable, and all the others follow after it in an immutable order.

E 5. A Digression About Notation

The symbol for a transition between two stationary states, and for the spectrum line which corresponds to that transition, consists of the symbols for the two states separated by an arrow, or a dash, or a semicolon, or any convenient mark. The final state is commonly written first. Thus the line due to the transition from a state B to a state A is designated thus: $(A)-(B)$. Chess-players will be reminded of the "Continental" system of describing moves at chess, in which symbols for the squares from which and to which the piece is moved are written down one before the other.

The notation for spectrum lines thus flows easily and naturally from the notation for stationary states. This notation is not in principle very difficult, but it has become confused and confusing, largely because of the alterations which have been wrought upon it to make it express not the facts, but divers theoretical interpretations of the facts. Alterations in names and notations generally produce an evil effect in physics even when justified in the highest degree, for the old systems and the new persist side by side and cause interminable trouble; all the more is this so when the alterations are based on uncertain grounds and impermanent. The notation for stationary states has already suffered much in this manner, and probably the worst is yet to come.

The classification of levels which I have just described enables and requires us to give a twofold symbol to each level; the symbol must designate the column in which the level stands, and its order-number or serial number in that column. The columns are generally desig-

[17] Foote, Meggers and Mohler observed a line corresponding to a change of two units in k (the line $(1,s)-(3,d)$, in the notation to be explained in section E5) under circumstances in which it seemed impossible to believe in an abnormally large electric field.

nated by the letters s, p, d, f (or their capital, or Gothic, or Greek equivalents).[18] A spectroscopist using these symbols generally writes the serial number of the level before the letter, with a comma between, thus: $(1,s)$ and $(2,p)$ and $(3,d)$. Or the columns may be designated by their values of the numeral k, which is then commonly written as a subscript to the serial number. These symbols have at least the advantage of being comparatively fixed. It is far otherwise with the serial numbers. One might expect that the level having the greatest energy-value in a particular column would be called Number 1, and the successive ones Number 2, Number 3, and so forth towards the convergence-limit. Unfortunately (though for not a bad reason) the habit is to designate the first levels of the successive columns by the order-numbers 1, 2, 3 and 4, successively; so that their respective symbols are $(1,s)$; $(2,p)$; $(3,d)$ and $(4,f)$. These are the symbols I have affixed in Fig. 6; but they are not the only ones, as the order-numbers have jumped up and down several times to satisfy the exigencies of new atom-models. It would be unprofitable to confuse the reader with further details, at least at this point. The important things to remember are three: that the symbol for each stationary state must contain one index for its column and another for its place in its column—that the former index is usually one of the specified letters—that the latter index is a number, usually beginning with 1, 2, 3, 4 for the first level in the s, p, d, f columns, respectively, and ascending along the column in unit steps.

E 6. Names and Features of the Most Noted Line-Series

Every line in every series, according to Bohr's fundamental idea, corresponds to a transition or "combination" between two stationary states of the atom—to a transition from an initial state to a final state. The atom possesses more energy in the initial state than in the final state (we are speaking of emission-spectra only). Hence the energy-value of the initial state, reckoned as it usually is from the energy of the ionized atom as zero, is algebraically higher and arithmetically lower than the energy-value of the final state.

The various lines of any one line-series have this in common: they correspond to transitions from various initial states which however all lie in one and the same column, into one final state which is the same for all and lies in an adjacent column. Each line-series thus

[18] The symbol b is sometimes used instead of f. For the columns following to the right of the f-column there are various notations, particularly f', f'', f''' and g, h, i. See also footnote 21.

belongs to one particular final state, and to one particular column of initial states.

The line-series consisting of transitions into the state $(1,s)$, or *terminating upon* $(1,s)$ as the phrase sometimes is, bears the name of *principal series*. Its consecutive lines are: $(1,s)-(2,p)$; $(1,s)-(3,p)$; $(1,s)-(4,p)$ and so forth. They are signified by the blue arrows of

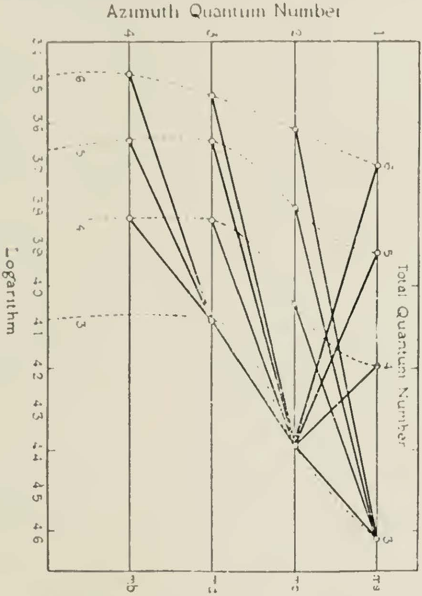


Fig. 7 Another way of mapping the stationary states of sodium

Fig. 6. The general symbol for this series is $(1,s)-(m,p)$; which will be quite intelligible. The $(1,s)$ level is the normal state of the atom; consequently, the various lines of the principal series correspond to transitions, by which the atom regains its normal state after a temporary exile from it. It is probably for this reason that the series is prominent enough to have received the name *principal* from the spectroscopists.

Two series terminate upon the $(2,p)$ level. One of these consists of transitions from various levels of the s-column. This is the *sharp* (or second) *subordinate* series, and its symbol is $(2,p)-(m,s)$. The other series consists of transitions from various levels of the d-column; it is the *diffuse* (or first) *subordinate* series, and its symbol is $(2,p)-(m,d)$. Yellow and red arrows signify these series, respectively, in Fig. 6. Of the two line-series terminating upon the $(3,d)$ level, only one has been endowed with a name; this is the series $(3,d)-(m,f)$, known alternatively as the *Bergmann* or the *fundamental* series (the second name is a bad one) and symbolized by green arrows in Fig. 7.

These series seem to be the only ones which impressed themselves strongly enough upon the minds of spectroscopic experts to receive names [19] from them. However, many other series have been identified, and emphasized, especially since Bohr's manner of thinking took root among the students of spectra; for instance, series terminating upon $(2,s)$ and $(3,s)$, which are conspicuous in the spectrum of helium, and such line-series as $(3,d)-(m,p)$, and $(4,f)-(m,d)$.

Several rules about line-series, which are very prominent in accounts spectra, become self-evident when the rules governing the stationary states are mastered (of course, this is only because the latter rules are based upon the former). For instance, there is a rule that the sharp and the diffuse series have the same limiting-frequency; and there is a rule that the difference between this limiting-frequency and the limiting-frequency of the principal series is equal to the frequency of the first line of the principal series. The reader may derive these by inspecting Fig. 6.

Such rules do not apply to the spectra of hydrogen and of ionized-helium, which are profoundly different from the spectra of sodium and other elements; and it is perilous to attach such names as *principal* or *subordinate* to the line-series of those first elements. The stationary states of those elements are known by their energy-values, and the series by the names of their discoverers or interpreters.

E 7. *Further Analysis of the Stationary States of Hydrogen and Ionized Helium; Fine Structure*

In our earlier analysis of the spectrum of hydrogen and the spectrum of ionized helium, we inferred from each of these spectra a family of stationary states, the energy-values of which follow one upon the other in a very regular procession governed by a simple numerical law. This makes it practically impossible to divide up these stationary states into classes; all of the levels for each of the atoms must

[19] The reader will recognize, in the initials of these names, the letters s, p, d, b, and f used to designate the several columns of levels.

inevitably be arranged in a single column, as it was done in Fig. 5. But in this arrangement the selection-principle of the foregoing paragraph is apparently contravened. For, when the levels of the sodium atom were arranged into columns, the transitions between levels belonging to one and the same column were among the inhibited transitions, the lines corresponding to these were among the missing lines. But the transitions between the levels in the single column which contains all of them for the hydrogen atom, correspond to the actual lines which constitute the entire hydrogen spectrum.

This discord is only apparent. It vanishes when we recall the fact, already once mentioned as a forewarning and then neglected for ease of exposition, that the stationary states of the hydrogen atoms are compound—that what has been called a "stationary state" in the preceding pages is really an ensemble of adjacent stationary states. Every line of the Balmer series, the series $R(1/m^2-1/2^2)$, is actually a close doublet; the frequency-differences between the components of all the doublets are approximately the same. Interpreted in the new fashion, this means that what we have called the stationary state of energy $-Rh/4$ is actually a pair of "component" stationary states very close together—so close together, that if the energy of one were exactly $-Rh/4$, the energy of the other would depart from that value by less than one part in forty thousand. Further in analyzing the spectrum of hydrogen we cannot go, probably because the minute details (if there are any) of the structure of its lines overtax the resolving-power of our spectroscopes. The spectrum of ionized helium, however, is spread out in a more generous scale; and some of its lines were analyzed by Paschen. Among these were the lines of frequency $4R(1/3^2-1/4^2)$; $4R(1/3^2-1/5^2)$; and $4R(1/3^2-1/6^2)$. They were resolved respectively, into six, five, and three components; and the line $4R(1/4^2-1/5^2)$ resolved into four.

Interpreted in the new manner, these data mean that what we have called the stationary states of energy-values $-4Rh/9$, $-4Rh/16$, $-4Rh/25$, and $-Rh/36$, are really ensembles of "component" stationary states lying very closely together. It would scarcely be possible to infer from these data, independently and without extraneous guidance, just how many "components" belong to each of the four ensembles. Fortunately or unfortunately, Paschen's measurements were preceded and inspired by a specific prediction of the number of components in each ensemble—a prediction that what we have called the nth stationary state should be a group of n "component" stationary states. This prediction is graphically set forth in the second column of Fig. 8, in which the level of energy-value

$-4Rh$ is drawn as a single dash, the next level as two dashes, the next as three, and so forth. Paschen's data were therefore compared with this prediction.

The data and the prediction were found compatible. If arrows are drawn from every "component" stationary state to every other "component" stationary state, it is found that each of the lines which was observed corresponds to one of the arrows (but it is necessary to assume that, in some places, two or more adjacent lines are fused apparently into one by reason of the insufficient resolving-power of the spectroscope). Some of the arrows, however, correspond to missing lines. Evidently some sort of inhibiting agency is at work; some sort of a selection-principle is adumbrated. Furthermore, some and perhaps all of the missing lines appear when the electric field strength acting upon the radiating atoms is increased, and this, it will be remembered, is the behavior of the missing lines in the sodium spectrum. Whether the selection-principle could ever have been inferred from these data alone seems doubtful. Naturally one proceeds to try out the same principle as served for the previous case. Can the component stationary states of the ionized-helium atom be sorted out into parallel columns, in such a manner that transitions between levels in adjacent columns correspond to actual, all the other transitions to missing, lines?

This is attempted in the manner shown in Fig. 8. The result is fairly satisfactory. The lines due to transitions between levels in adjacent columns should by this principle be visible, and they are. The lines corresponding to transitions between levels in the same column, or more than one column apart, should be missing; and some of them are, but also some of them undeniably can be seen. To account for these unwelcome guests, it is necessary to assume that some of the radiating atoms are subject to a strong electric field which might, but would not be likely to, exist in the discharge. This is an uncomfortable solution; but there are other numerical agreements between the prediction and the data, which it is not expedient to describe at this point, but which are good enough to excuse that deficiency to some extent. *En somme*, the evidence presents no insuperable objection to our arranging the component stationary states of the ionized-helium atom in parallel columns, and declaring that the only transitions which occur (except in strong electric fields) are those between members of adjacent columns; and this is just what we did with the sodium atom, and can in general do with every other kind of atom whereof the spectrum has been interpreted. This being granted, we can assert that the spectra and the stationary states of the ionized-helium atom (and presumably those of the hydrogen atom) are not so radically different from those of the sodium atom as they seemed to be; some of the apparent differences being traceable to the fact that corresponding levels in the *f*, the *d*, the *p*

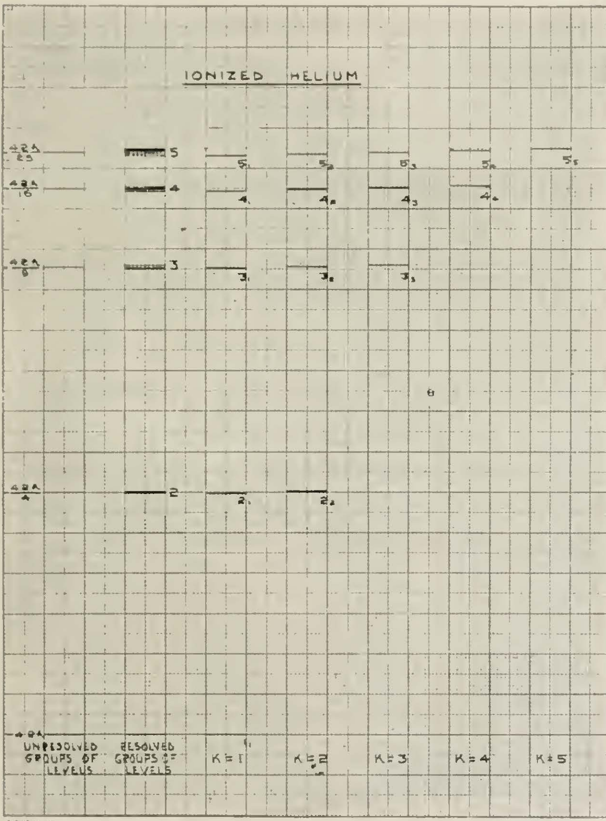


Fig. 8—Diagram of the stationary states of ionized helium, resolved to account for the fine structure of the spectrum lines

and the *s*-columns, which in the sodium atom are widely separated, are in the former atoms so closely crowded together that lines, which in the sodium spectrum are far apart, are in the former spectra packed into all-but-irresoluble groups. This is probable, but not certain. Further data about other lines in the ionized-helium spectrum would be gratefully received.[20]

The notation for the various "component" stationary states of the ionized-helium atom is shown in Fig. 8. The successive columns are denoted by the numerals 1, 2, 3, 4 . . . for which the general symbol is k, as previously. This numeral is written as a subscript to the serial number of the level in its column, which commences with 1 in the first column, 2 in the second, 3 in the third, and so forth. By inspecting the figure, the reader will see a reason for using these different values of the serial-number for the first levels of the different columns. The serial-number is designated by n and called the *total-quantum-number*. The numeral k is called the *azimuthal-quantum number*, as before. These heavily long names are imposed by the theory and not by the data.

E 8. *Further Analysis of the Stationary States of Other Elements than Hydrogen and Ionized Helium*; *Multiplets*

Having performed a two-stage analysis of the spectra of ionized helium and of hydrogen, we return to the spectra of the other elements for a second attack.

Let us consider the reasons for making these analyses in two stages. When the mid-Victorian physicist trained his spectroscope upon a tube full of glowing hydrogen, he saw the spectacle of Fig. 2—the converging procession of distinct bright lines, of which the frequencies form that delightfully smooth numerical progression which we have already met. Later physicists with better instruments discovered that each of these "lines" was in fact a pair of lines. Now in strict truth, this discovery showed that the "lines" of the Balmer series were no lines at all; for a doublet is not a line. But the physicists continued to refer to the "lines" of the Balmer series, chiefly no doubt because to anyone equipped with an ordinary spectroscope the doublets do appear as single lines. By itself this is little reason; but the usage is not altogether faulty. Few people would hesitate to admit that each of these doublets is not a couple of casual neighbors, not two

[20] It would be particularly interesting to settle beyond question whether the missing lines demand the selection-principle already explained in section E4, rather than the one to be explained in section E8. This is one of the reasons for wanting to produce and examine the spectrum of doubly-ionized lithium, in which the evidence would probably be much clearer.

unrelated lines fortuitously close together, but a pair of lines sharing some deeply fundamental quality in common. This is indicated chiefly by the facts that the distance (measured in frequency) between the components of a doublet is the same for all the doublets, and very small compared with the distance between consecutive doublets. For this reason the doublets are treated as entities, and they require a name; which is what physicists have preserved for them, in continuing to call them "lines." "Doublet" would be better than "line", and "group" would be better yet; but we cannot ever be sure that even the apparently-single lines are not very close groups, and yet it would be silly to call every line a group. Sirius appears as a double star in a few of the most powerful telescopes, but nobody would insist on calling it a double star when pointing it out in the night sky.

All this is not so trivial as it sounds. It is easy enough to speak of doublets when looking at lines which appear single except when viewed in the most powerful spectroscope, and then are resolved into components much closer together than the nearest similar line is to either. Such lines occur not in the spectra of hydrogen and ionized helium only, but in the spectra of sodium and other elements generally. But the spectroscopist is constantly applying such names as "doublet" and "triplet" and "quadruplet", and the inclusive name "multiplet" to groups of lines which lie far apart in the spectrum, with scores of others intervening. Here his function is not to split apparent lines into narrow groups, but to unite widely-scattered lines into wide groups. This he does not because of propinquity of the lines, but because of resemblances or analogies or fixed intensity-relations between them, or because he finds it possible to construct a series of such groups with identical frequency-differences between corresponding lines within them, or because of analogies with other elements with more perspicuous spectra, or theoretical predictions, or intuitions or clairvoyance. Groups such as these are not generally termed lines, except in very abstract discussions; it is difficult to call a group a line, when it is clearly resolved by any instrument worthy the name of spectroscope. But they are like the lines of the Balmer series, treated as entities because their lines are believed to share some deeply fundamental quality in common.

What I have said about lines and groups of lines is transferable in substance to stationary states and groups of stationary states. What we had originally called the levels of hydrogen and ionized helium, with their energy-values $-Rh/n^2$ and $-4Rh/n^2$ ($n = 1, 2, 3 \ldots$), were resolved into groups of levels in order to interpret the fine structure of the lines. But owing to the propinquity and to certain numerical relations of the levels in a group, and to certain qualities of the transitions between them, it was felt that the levels of each group share some deeply fundamental quality in common. For this reason we used a system of classification in which each level is represented by two symbols, one for its group and one for its place in its group; and we numbered the levels in succession, not 1 and 2 and 3 and 4 and 5 and so forth, but 1_1 and 2_1 and 2_2 and 3_1 and 3_2 and 3_3 and so forth. Interpreting the groups of lines in the spectra of sodium and other atoms, we infer groups of levels. The levels in one of these groups are often far apart. They may be eighteen or more in number, other levels may lie between; but by reason of the resemblances between the lines whence they were inferred, by reason of certain numerical relations between the levels themselves, they are believed to have some deeply fundamental quality in common. If this is vague, so also at times is the interpretation.

The statements in the foregoing sections about the stationary states of sodium are now to be understood as relating to groups of stationary states. It is the *groups of stationary states which are arranged in parallel columns, designated by numerals* k, *such that no transition takes place unless in it* k *changes by one unit.* It is the group of stationary states which is marked by a pair of numerals, one to designate its column and the other its place in its column; or by a letter to designate its column and a numeral to designate its place in its column. It is the group of stationary states which is denoted by (3_2) or $(1,s)$ or $(5,d)$.

To denote a particular stationary state we must add, to the symbols for its group, a third symbol for its place in its group. This symbol is generally a numeral, hung on as a subscript to the letter designating the column (thus: $(2,p_1)$ and $(2,p_2)$) or as an additional subscript to the two numerals (thus: 3_{21} and 3_{22}).[21] The most common general symbol for this numeral is j. Geometrically, the stationary states may be represented by lines or dots arranged, not in one row of several parallel columns as in Fig. 7, but in several rows of parallel columns. Readers with three-dimensional imaginations in good working order may develop this idea *ad libitum*. The systems for assigning the values of j are shifted around every few months to correspond to new atom-models, and are scarcely worth memorizing.

[21] The notation suggested by Saunders and Russell, evidently in concord with a number of other experts, is built in this way: Designate the column to which a group belongs by the letters suggested in section E5, capitalized (i.e., S, P, D, F, G, H for $k = 1, 2, 3, 4, 5, 6$); write the serial-number of the group before the letter, and append the value of j as a subscript to the letter. If it is desired to state what sort of a system (cf. section E10) a level belongs to, one may add an index to the left of the letter and above it.

The best of them, however, are adjusted so as to express a new and additional selection-principle, which is coequal with the other selection-principle we met a few pages above.

This principle is derived in the same way as the first one. The groups of levels are established by inference from the groups of lines; then arrows are drawn from every level to every other, the corresponding spectrum-lines are sought, and most of them are not found. Some of these missing lines are those which would contravene the first selection-principle, as they correspond to transitions in which the numeral k changes by more than one unit, or not at all. Putting these aside, there are still a number of missing lines, to which the first selection-principle has offered no objection. Now it is found possible to choose the numeral j in such a manner that the only transitions which correspond to actual spectrum lines are those in which j changes by one unit or not at all ($\Delta j = O, \pm 1$). Furthermore it is possible to adjust the values of j in such a manner that the lines corresponding to transitions, in which j is initially zero and remains unchanged, are missing.

This is the *selection-principle for the inner quantum number*; for the numeral j, when adjusted in this manner, is known as the inner quantum number. This again is a name imposed by theory and not by the data of experience.

As the two selection-principles are effective concurrently, the pair of them may be fused into this one:

Of the three numerals n, k *and* j, *which specify a stationary state completely, two* (k *and* j) *may be so chosen that the only transitions which correspond to actual lines are those in which*: *first*, $\Delta k = \pm 1$; *second*, $\Delta j = O, \pm 1$; *third*, j *is not zero both before and after the transition.*

This complicated rule is evidently the sign of some very important principle, the full nature of which thus far escapes us. It will probably seem difficult to grasp and fix in mind; but difficulty of this sort is likely to abound in the physics of the near future. Not so many years ago the physicist's path lay among differential equations; the defter he was in integrating hard specimens of these, the better he was fitted for his profession. I should not care to say that this is no longer true; but he will probably have to cultivate a sense for problems such as this.

It remains to give some idea about the number of stationary states in the various groups. For sodium, as laid out in Fig. 6, the groups in the *s*-column are merely single levels (this sounds like a contradiction in terms, but may be borne for the sake of the generality); the groups in the other columns are pairs of levels, or "doublet terms."

This is the common character of the alkali elements Li, Na, K, Rb and Cs, which occupy the first column of the periodic table; probably also for the noble metals which share this column, but the data are few. For elements of the second column of the periodic table there are two complete systems of stationary states, each having its own *s*-column, its own *p*-column, its own *d*-column, and all the rest. In one system, all the groups in every column reduce to single levels; it is a *singlet system*; in the other, all the groups in the *s*-column are single levels, all the groups in the other column are triads of levels or "triplet terms;" it is a "triplet system." The complexity mounts up stage by stage as we cross the periodic table of the elements from left to right, and soon becomes terrific.

E 9. Effect of Magnetic Field on the Stationary States

When a magnetic field is applied to a radiating gas, most of the lines of its spectrum are replaced by triplets (Fig. 9), or by even richer groups of lines (Fig. 10). By a somewhat loose usage the lines are said to be *resolved* into three or more components. This is the "Zeeman effect." There is a multitude of empirical rules about these components, their spacings, the way in which their number and their spacings vary from one line to another, and other features. According to the new fashion, however, we focus our attention not on the component lines, but on the stationary states which are inferred from them.

The effect of a magnetic field may be described by saying that it replaces each stationary state (with a few exceptions) by two or more new ones. Each of these new states requires four symbols to designate it; the symbols n, k and j for the original stationary state, and a new symbol m to denote its place in the resulting group. As heretofore, when every stationary state is connected with every other by an arrow and the corresponding lines are sought, it is found that some of the lines are missing. Still another selection principle is therefore to be sought, and the values of the new numeral m are to be so adjusted—if possible—that the selection-principle can be read easily from them. When so adjusted m is called the *magnetic quantum-number.*

In certain cases the empirical rules for the components whereby the magnetic field replaces the individual lines are simple; and the derived rules for the new stationary states which arise out of the original ones when the magnetic field is applied are correspondingly simple. These are the cases of "normal" Zeeman effect (the adjective "normal" may be an entirely misleading choice). Let ΔU_m

represent the energy-difference between the new stationary state denoted by the index *m*, and the original stationary state. The rules are comprised in the formula,

$$\Delta W_m = m\omega Hh \tag{8}$$

and in the selection-principle. In the formula H stands for the magnetic field, ω is a factor equal within experimental error to $e/4\pi\mu c$ (μ = mass of the electron) and commonly identified with it. m has two or more values spaced one unit apart (for instance, 1 and 0, or $\frac{1}{2}$ and $-\frac{1}{2}$, or 1 and 0 and -1).

The selection principle is as follows: *The only transitions which correspond to actual lines are those in which* m *changes by unity or not*

Fig. 9 Effect of magnetic field on spectrum lines. (P. Zeeman, *Journal of the Franklin Institute*)

at all: $\Delta m = 0, \pm 1$. This is the *selection-principle for the magnetic quantum number.*

If we allow m to assume only two values, this principle becomes nugatory. If on the other hand, we adopt the principle, m can assume any number of values whatever, provided only they are spaced at unit

Fig. 10—More complicated effects of magnetic fields on spectrum lines. P. Zeeman, *l.c.*

intervals; it makes no difference with the observed lines whether there are two or two hundred new stationary states for every original one. This is convenient for theorizing. In dealing with the Zeeman effect in general, and not merely with these special "normal" cases, it is necessary to assume that ω is not restricted to the particular value just given, but depends on the stationary state in question; and that *m* depends on the value of *j* for the stationary state in question.

Very strong magnetic fields treat a group of stationary states as if they were one single state—as if they were first all fused together into one, and this one then resolved according to equation (8). This is the *Paschen-Back effect*. It evidently means a great deal.

The light emitted from a gas exposed to a magnetic field is polarized. Some of the new lines are circularly polarized about the direction of the magnetic field as axis; others are plane-polarized, with the electric vector parallel to the direction of the magnetic field. The lines corresponding to transitions in which *m* changes by one unit are all polarized in the former way; the lines corresponding to transitions in which *m* does not change are all polarized in the latter way.[22]

E 10. *Interrelations of Multiplets and Zeeman Effect*

I insert this section chiefly for the benefit of such readers as may be preparing for a thoroughgoing study of atomic theory. Others may do well to pass it over, as the statements it contains can scarcely be apprehended with any vividness, except by the aid of pencil and paper and hours of reiteration. For those who omit this section I will merely say, that the material described in it goes far to show that the numerical values which we have been assigning to *k* and *j* are not quite arbitrary, but are determined by something fundamental; although the ones heretofore assigned are not necessarily the most expressive.

I begin with a description of the various known systems of stationary states, condensed into Table III. To make this table clear I will explain the fourth line; this line contains the statement that a "quartet system" of stationary states consists of an *s*-column of single levels, a *p*-column of groups of three levels each, and a *d*-column, an *f*-column, and additional columns of groups of four levels each.

TABLE III

Name of System	*s*	*p*	*d*	*f*	*f'*	*f''*
Singlet	1	1	1	1	1	1
Doublet	1	2	2	2	2	2
Triplet	1	3	3	3	3	3
Quartet	1	3	4	4	4	4
Quintet	1	3	5	5	5	5
Sextet	1	3	5	6	6	6
Septet	1	3	5	7	7	7
Octet	1	3	5	7	8	8

[22] The effect of a magnetic field on resonance-radiation, discovered by Wood and Ellett, will be described in the Second Part.

Elements of the first column of the periodic table possess a doublet system of stationary states; elements of the third column, a doublet system and in addition a quartet system. It is inferred that elements of the fifth column possess these and a sextet system in addition; elements of the seventh, these three and an octet system in addition. Elements of the second column of the periodic table possess a singlet system and in addition a triplet system. It is inferred that elements of the fourth column possess these two and a quintet system in addition; elements of the sixth column, these three and a septet system; elements of the eighth, these four and a nonet system. These inferences have been partially verified. For titanium, in the fourth column of the periodic table, the triplet and quintet systems have been discovered; for vanadium (fifth column) the quartet and sextet; for chromium (sixth) the quintet and septet; for manganese (seventh) the quartet, sextet, and octet; for iron (eighth) the triplet, quintet, and septet. Apparently it is by no means certain that the unmentioned systems are really missing, as the difficulties of analyzing these complex spectra are terrific.

There are certain rules governing the number of levels in a group, and the effect of a magnetic field upon these levels. These rules were discovered chiefly by Landé; I give them in his notation. I recall, to begin, that we have designated each group of levels by a numeral k, which is 1 for all the groups in the s-column, 2 for all groups in the p-column, 3 for all groups in the d-column, and so forth. We have further distinguished the different levels in a group by assigning them different values of another numeral j; the manner in which these values of j are chosen was described in section E8. Landé introduces a numeral K which is smaller than k by $\frac{1}{2}$; K thus is $\frac{1}{2}$ for all groups in the s-column, $1\frac{1}{2}$ for all groups in the p-column, and so forth. He also introduces a numeral J which is greater than j by $\frac{1}{2}$; and a numeral R which is $\frac{1}{2}$ for every level belonging to a singlet system, $2/2$ for every level belonging to a doublet system, $3/2$ for every level belonging to a triplet system, and so forth.

These are Landé's rules:

(1) The total number of levels in a group characterized by the numeral K, belonging to a system characterized by the numeral R, is twice the smaller of the two numerals R and K (that is, it is $2R$ if $R<K$; $2K$ if $R>K$; $2R=2K$ if $R=K$).

(2) In the formula (8) for the Zeeman effect, the factor ω is equal to $e/4\pi\mu c$ multiplied by a factor g, which depends on the numerals R, K, and J for the level in question in the following manner:

$$g=3/2+(R^2-K^2)/2(J^2-\tfrac{1}{4}) \qquad (9)$$

(3) In the same formula, the magnetic quantum-number m depends on the numeral J for the level in question; it assumes $2J$ values altogether, commencing at the maximum value $(J-\frac{1}{2})$ and going downwards across zero to $(-J+\frac{1}{2})$.

These rules form a beautiful little problem for the designer of atom-models. They have often been tested and verified (it is not easy to find out just how far), and at present are widely used in the deciphering of spectra. It appears, however, that some spectra—particularly those of the inert gases—are too complicated even for these rules, and possess a structure even more elaborate. Considering how difficult it is to grasp the structures already described, one may be excused for feeling some dismay at the prospect.

E 11. Effect of Electric Field on the Stationary States

When an electric field is applied to a radiating gas, the lines of its spectrum are replaced by groups of lines, often rich and complicated.

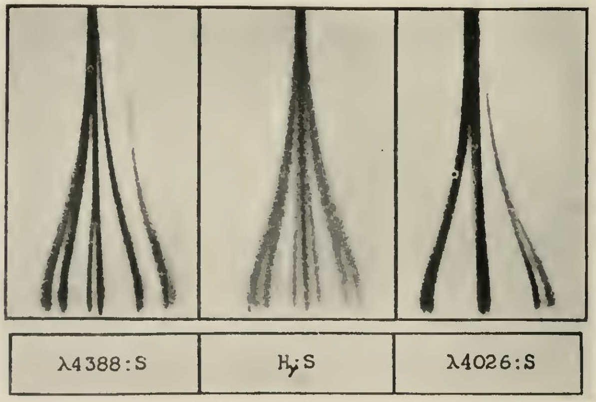


Fig. 11—Resolution of spectrum-lines into groups, displacement of lines, and emergence of missing lines, produced by a strong electric field (increasing from the top downwards to nearly the bottom of the picture). (J. S. Foster, *Physical Review*)

(Fig. 11.) From these we infer, as heretofore, that the stationary states are replaced by groups of stationary states. The atom-model proposed for hydrogen and ionized helium has been extraordinarily successful in describing the effect of electric field upon their spectra, and therefore I shall violate the rule I have heretofore followed, and postpone the description of the phenomena until the theory is stated. Atoms of other kinds are affected in at least two ways; the stationary states are displaced, and the "missing lines" are evoked, as I have said already.

E 12. Intensity-Ratios

The relative intensities of the various lines of a doublet, or triplet, or multiplet are often equal within the (fairly large) uncertainties of measurement to simple ratios, such as 1 : 2, 2 : 3, 3 : 4. This happens too often to be easily put down as a mere coincidence, and indicates that the occurrence of transitions is governed by simple laws. Our selection-principles are themselves indications of the same type, since they may be taken as signifying that the intensity-ratio of certain lines to certain others is zero. This problem may be more difficult than the ones I have stressed hitherto, since each line involves two stationary states and is not a quality of one only. This applies to other properties of lines, such as their sharpness or diffuseness.

E 13. Excitation of Individual Frequencies

So long as an atom is conceived as a belfry full of bells of various pitches, it would probably be argued that a shock to the atom would set all the bells to jangling, and a gas bombarded by electrons would emit all of its natural frequencies if any. The interpretation of spectra to which these pages are devoted leads to a very different idea. A spectrum-line of frequency ν is emitted when the atom passes from a stationary state B to a stationary state A. The energy-value of state B by itself does not determine ν; this is controlled by the difference between the energy-values of B and A, which is $h\nu$. But the energy-value of B has everything to do with whether or not the frequency ν is emitted under given conditions; for it will not be emitted at all unless the atom is first put into state B. If the gas is bombarded with electrons of energy insufficient to raise an atom from its normal state to state B, then the line in question, and all of the other lines which result from transitions from B to other levels of lower energy-value, will fail to appear. If the energy of the electrons is raised past the critical value (the difference between the energy-value of B and the energy-value of the normal state) all of these lines suddenly appear.

This is illustrated by Fig. 12, relating to magnesium. An electron striking a magensium atom and having an energy equal to 3.2 equivalent volts is able to put the atom into a particular excited state; the atom emits radiation of wavelength 4571 in returning to its normal state. To get the atom to emit another sort of radiation, the electron must possess 6.5 equivalent volts to put it into another excited state. Any excited state can be reached if the electron has 10 equivalent volts to pass over to the atom.

In a gas sustaining an electrical discharge, the atoms are subject to stimuli of such variegated force and type that the distinctions between different lines are not so clearly marked; but it can be seen

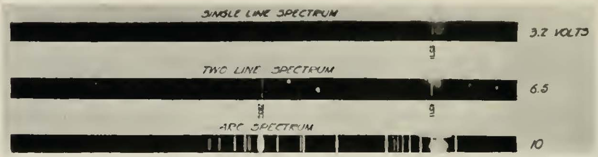


Fig. 12—Successive excitation of lines requiring electron-impacts of successively greater violence to bring atoms into the necessary initial states. (Foote, Meggers, and Mohler, *Philosophical Magazine*)

that mild discharges favor lines for which the initial level is adjacent or close to the normal level, while other lines require a more violent stimulus. Furthermore, when a gas is steadily heated to higher and higher temperatures, various lines of its spectrum appear in more or less the order of the stationary states which are the initial states of the transitions responsible for these lines. Accordingly a "temperature classification" of spectrum lines has been developed at Mount Wilson Observatory and elsewhere, and is valuable in deciphering intricate spectra.

E 14. Absorption-Spectra

An atom which will emit a frequency ν when it is originally in a state B and passes over into a state A, will absorb light of the same frequency if it is initially in the state A. This has the important consequence that the lines which a gas absorbs, when lying at rest and unexcited, are those which it emits in passing from any and every other state *into the normal state.* The lines emitted when an atom passes from one of its stationary states into another which latter is not the normal state, are not absorbed by the gas lying quiescent and undisturbed. For this reason helium and neon and argon are quite transparent to all visible light, although they have many emission-lines in this region of the spectrum; for each of these lines corresponds to a transition into some other than the normal state and the lines which correspond to transitions into the normal state lie far off in the ultraviolet. But if such a gas is made the theatre of a self-sustaining electrical discharge, the other lines likewise are absorbed: for the discharge puts the atoms of the gas temporarily but frequently into various abnormal states. This incidentally is one of the bits of evidence that an atom may sojourn for a finitely long time in another stationary state than the normal one. If the gas is heated, the same effect occurs; for the violent collisions between atoms in a hot gas occasionally bring atoms into excited states.

By observing the absorption-spectrum of a quiescent gas one learns which lines in the emission-spectrum correspond to transitions into the normal state a valuable piece of information in the cases of elements of which the spectra are complicated and obscure.

E 15. Spectra of Ionized Atoms

In a violent electrical discharge, such as a spark, the gas emits many lines which cannot be fitted into the system of series of the usual spectrum of the gas. These may also be produced by bombarding the gas with electrons possessing more than enough energy to ionize its atoms. They are believed to emanate from ionized atoms, or from atoms deprived of one electron. The spectrum of ionized-helium has been very important in these pages. In very violent sparks many more lines emerge, and these are associated with atoms deprived of two, three, or even more electrons.

The spectrum of the ionized atom of an element resembles, in its system of series, and in more minute details, the spectrum of the neutral atom of the element preceding it in the periodic system. The spectrum of an atom deprived of n electrons resembles the spectrum of the neutral atom preceding it by n places in the periodic system. This confirms the belief that the spectrum and the other properties of an element are determined chiefly by the number of electrons which its atom contains.

E 16. X-ray Spectra

The difference between the X-ray spectra to which we now come, and the "optical" spectra which we have been discussing seemed profound and vital in the era of very defective knowledge, but it has faded steadily away with the deepening of understanding. Twelve or fifteen years ago the contrast was multiform and very sharp; for the optical spectra were produced chiefly by maintaining an electrical discharge in a gas, the X-ray spectra invariably by bombarding a solid body with exceedingly fast-moving electrons or with other X-rays; the optical frequencies could be diffracted and refracted, the X-rays not at all or almost imperceptibly little; the optical frequencies were all inferior to 3.10^{15}, the X-ray frequencies all clearly more than a thousand times as great. Since then, rays of almost all the intermediate frequencies and with intermediate properties have been generated in a variety of ways, and the distinction is no longer trenchant, except between the extremes. To make it so, one must seek a theoretical reason—and perhaps there is none to be found.

There is, however, apparently good ground for introducing a theoretical distinction. I have pointed out heretofore that the energy which an atom loses, when it radiates one of the lines of its "optical" spectrum, is less than the ionizing-energy. Or, turning this statement around and amplifying it a little: the energy which an atom absorbs, when it absorbs one of the rays of its optical spectrum, is less than what is required to detach the loosest electron from it. Therefore it is possible to assume, at least as a trial hypothesis, that the energy is spent in lifting the loosest electron partway out—a hypothesis fortified by the fact that, when the atom has just absorbed some energy in this manner, the electron can be detached by supplying the atom with enough extra energy to bring the total amount up to the ionizing-energy. But if we take one of the typical X-ray frequencies, and multiply it by h to ascertain how much energy the atom gains in the process of absorbing that frequency, we find that the quantity $h\nu$ exceeds the ionizing-energy tremendously. This circumstance makes it quite out of the question to imagine that the X-rays are due to changes in the position or the motion of the loosest electron alone. We may therefore define the X-ray frequencies as those which cannot be explained as due to transitions of the loosest electron, from one motion or position to another, unaccompanied by other changes. By this definition, every frequency ν for which the quantum-energy $h\nu$ is greater than the ionizing-energy, goes into the X-ray spectrum. For the remaining frequencies the question is more dubious, perhaps never quite to be settled unless and until complete theoretical classification of all the lines is attained. In this section, however, I shall speak only of frequencies hundreds or thousands of times greater than the ionizing-frequency.

Gazing upon typical X-ray emission spectra one sees that they consist of groups of lines with wide intervals between. Going from higher frequencies towards lower, the groups are known successively as the *K* group, the *L* group, the *M* group and the *N* group. The word *series* is more commonly used than *group*; but this is a misfortune, for it suggests a dangerously misleading analogy with the series in the optical spectra which we have studied with so much care.[23] The process of measuring these lines and classifying them

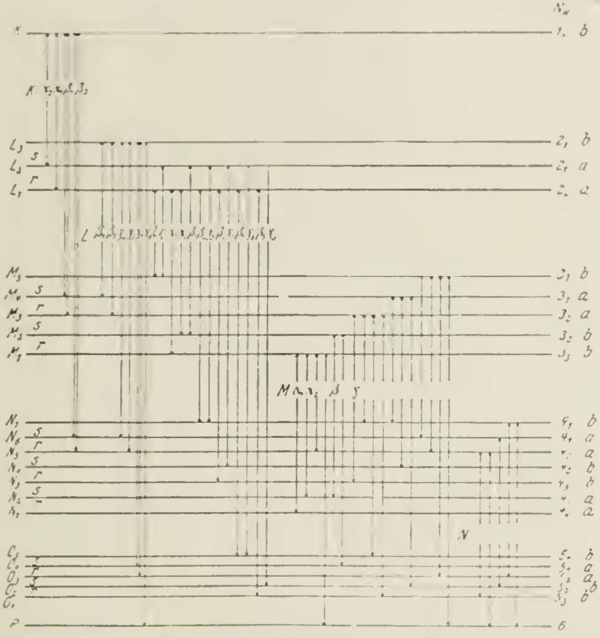


Fig. 13—Diagram of stationary states designed to account for the X-ray spectrum of uranium. (From Siegbahn, after Coster)

was carried out after the dissemination of Bohr's great idea that each line-frequency should be multiplied by *h* and the product interpreted as the difference between the energy-values of two stationary states of the atom. The complete analysis of an X-ray spectrum

[23] In fact the usage is inverted. A series, in the optical spectrum, is a set of lines having the same *final* state in common; but the "*k-series*" is a group of lines having the same *initial* state in common, the *L*-series a set of 3 groups corresponding to 3 initial states.

thus culminates in a diagram of stationary states, as for the optical spectra.

Such a diagram is shown in Fig. 13, which is for an element far up in the periodic system, therefore with a rich system of X-ray lines and stationary states. In comparing it with one of the diagrams made for optical spectra, it must be remembered that its scale is enormously more compressed—the distance from top to bottom corresponds to about one hundred thousand equivalent volts. Each line in the X-ray spectrum corresponds to an arrow between two of the levels, but not every arrow corresponds to a line. Again there is a selection-principle, and this selection-principle is partly expressed by attaching a double index to each of the levels. When the indices are assigned as in Fig. 13, transitions between levels for which the second numeral differs by one unit include the only ones which actually occur. But this is not the complete selection-principle; it is necessary to add that in any actually occurring transition, the first numeral must change by one or more units; and further, that transitions may occur only between levels to which different letters are attached. The first numeral is designated by n, the second by k; they are called the total and the azimuthal quantum-number.

The levels are also frequently known by letters with subscript numerals, as the diagram shows. The letters by now are pretty definitely fixed, but the subscripts are still being shuttled around. The notation for the X-ray lines is in a terrible state.

A curious and evidently important feature of these levels is, that when an atom is put into any one of them—say into the K level, or the L_1 level, or the L_2 level—it extrudes an electron. Or, in other words, each of these stationary states is a state in which the atom lacks one of its electrons—like the "ionized-atom" state from which we previously measured the energy-values in dealing with the optical spectra. All of them, at least the highest ones, are in fact "ionized-atom states." Since, however, they are all different, it is natural to suppose that a different electron is missing, or that an electron is missing from a different place, in each of the different cases. Apparently an atom cannot enter into a stationary state with so high an energy, and remain neutral.

We must pause to consider from what standard state the energy-values of these stationary states are measured. In the previous case of the optical spectra, the energy-values of the stationary states were measured, so to speak, *downwards* from the state of the ionized atom to the normal state of the neutral atom; the energy of the ionized atom was set equal to zero, that of the neutral atom in its normal state then had a certain negative value, all the other energy-values were negative and scattered between these two. In this case of the X-ray spectra, the energy-values of the stationary states are measured *upwards* from the normal state of the neutral atom, to which the energy-value zero is assigned, while all the other energies are positive. In Fig. 13 this zero-line must be imagined just under the level marked P.

The exact position of this zero-line for the high energy stationary states is not very accurately known; although the distance between any two levels is determined with all the usually very great exactitude of X-ray wavelength-measurements, the distance from any level to the zero-line is uncertain within a few tens of volts. This uncertainty is not great enough to be important when dealing with the high-frequency X-rays.

This point being attended to, we are now in position to consider the striking difference between X-ray emission-spectra and X-ray absorption spectra—striking indeed when one looks at typical photographs, apparently altogether a different matter from the contrast between optical emission-spectra and optical absorption-spectra, yet in principle very much the same thing. In dealing with optical spectra, I remarked that while an atom *may* absorb any frequency which it can emit—while the complete absorption-spectrum of a gas is identical with its complete emission-spectrum, yet the absorption-spectra one ordinarily sees contain only a small selection of the emission-lines. This occurs because when a gas is being examined for its absorption-spectrum in the laboratory, by sending light through it, it is generally in an untroubled and quiescent condition, each of its atoms being in the normal state; therefore it absorbs only such frequencies as provoke transitions from the normal state to the various excited states, and not such frequencies as would induce transitions from one excited state to another, for few or none of the atoms are in any one of the excited states to start with. Such also is the case with the X-ray spectra. Quiescent atoms absorb only such X-ray frequencies as produce transitions from the normal state into one of the stationary states designated by K, or L_1, or L_2, and so forth—they do not absorb such frequencies as would produce the transitions from L_1 or L_2 to K, for instance, for the atoms are not initially in the states L_1 or L_2. This is quite the same behavior as is observed in the response of atoms to radiations in their optical spectra. It is much more pronounced, however; for, while it is possible to make a gas absorb frequencies which produce transitions from one excited state to another, by maintaining the gas in a state of intense electrical excitation, this has never been done with metals or gases exposed to X-ray frequencies.

Atoms therefore do not absorb such X-ray frequencies as are represented by the downward-pointing arrows in Fig. 13. They do absorb such frequencies as would be represented by arrows drawn from the very bottom of the diagram—a little below the level marked P—up to the various levels; and (it may seem, unexpectedly) they also absorb frequencies somewhat higher than these. This however does not mean that the atom may be put into an excited state of higher energy than the K state, for instance; it means simply, as direct evidence proves, that the extruded electron receives the extra energy and goes away with it. Owing to this fact, the X-ray absorption-spectrum consists not of sharp absorption-lines at the several frequencies corresponding to a transfer of the atom into the K-state, the L_1-state, and so forth, but of continuous bands commencing with sharp edges at these frequencies, and trailing out gradually towards higher frequencies.

Another curious feature of the X-ray spectra is that transitions from the various excited states of high energy-values, such as the K-state and the L-states, directly into the normal state, apparently do not occur.

E 17. Band-spectra

Band-spectra are the spectra of molecules,—that is to say, of clusters of two or more atoms, such as appear in certain gases. This is proved by the fact that they are displayed by gases which are known in other ways (gramme-molecular volume, specific heat) to consist of molecules; by the fact that the band-spectrum of such a gas disappears when the gas is heated to the point where its molecules are dissociated into atoms; and by the general successfulness of the quantitative theory based on the assumption that they are due to molecules. Occasionally band-spectra are displayed by gases which are not otherwise known to contain molecules, such as helium and potassium; it is supposed that they are due to molecules too few to be detected by the other accepted methods. Usually they are easy to distinguish at first glance from the optical spectra of atoms, although there are exceptions, such as the band-spectrum of the hydrogen molecule. Like the spectra we have discussed, they consist of lines; the term "band-spectrum" describes the manner in which these lines are grouped. Again like the spectra we have discussed, they are analyzed according to Bohr's fundamental principle, by interpreting the lines as the results of transitions between stationary states.

F. Magnetic Moments of Atoms

Of the enormous and chaotic variety of facts about the magnetic properties of materials, only a few of the least conspicuous have been serviceable to atom-builders; the notorious ones have helped very little or not at all. The famous and characteristic magnetic properties of iron, nickel, cobalt, depend on the arrangement of the atoms and on the temperature of the metal, and cannot safely be attributed to the atoms themselves. Diamagnetism, an inconspicuous and rarely-mentioned quality of certain elements, is in some instances quite independent of temperature, and may well be a property of the atoms. Paramagnetism, an almost equally inconspicuous quality of certain other elements, depends on temperature, but in such a way that it may sometimes be explained by assuming that each atom has a characteristic magnetic moment, the same for all the atoms of a substance. The value of this magnetic moment of the atom may be calculated from measurements on the paramagnetism of the substance; the process of calculation involves certain assumptions, at least one of which is at the present open to question.

Direct measurements upon the magnetic moments of certain atoms are now being made by Gerlach; and they are among the most important achievements of these years. In a small electric oven, a metal such as silver is vaporized; a beam of the outflowing atoms, passing through a small orifice in the wall of the oven and through others beyond this one, eventually travels across a strong magnetic field with a strong field-gradient and falls upon a plate. Suppose that each atom is a bar-magnet, oriented with its length parallel to the magnetic field. If the field were uniform, the bar-magnet would not be deflected, it would travel across the field in a straight line; for although its north pole would be drawn sidewise by a force, its south pole would be pushed by an exactly equal force in the exactly opposite direction. That the atom may be drawn aside, the field must be perceptibly different at two points as close together as the two poles of the magnet. When one considers how small an object the atom is, it is clear that the field must change very rapidly from one point of space to another, its gradient must be enormous. Gerlach succeeded in contriving so great a magnetic field with so great a gradient that the beam of flying atoms was perceptibly drawn aside. The most-deflected atoms are those of which the magnetic axes are most nearly parallel to the magnetic field. From their deflections, the field, and the field-gradient, the magnetic moment of the atom can be computed very simply. The values thus obtained are of the order of 10^{-10} in *CGS* units.

I shall comment in the second part of this article upon other inferences from these experiments, which are as valuable as the experiments upon the transfer of energy from electrons to atoms. At this point it is sufficient to realize that these experiments prove that atoms, or at least the atoms of some elements, possess magnetic moment. If magnetic moment is due to electric current flowing in closed orbits, as Ampere and Weber guessed a century ago, the atom must be supposed to contain such currents; if the atom consists of a nucleus and electrons, some at least among the electrons must be supposed to circulate. And if the electrons are assumed to circulate in a particular manner the magnetic moment of the atom so designed can be computed, and thereupon tested by experiment.

This completes the list of the phenomena, the properties of matter, which are used in designing the contemporary atom-model. Nobody will require to be convinced that it is not a list of all properties of matter, nor of all phenomena. These are not among the obvious and familiar qualities of matter; and no one meets any of them in everyday life, nor perceives any of them with his unaided senses. They are phenomena of the laboratory, discovered after a long and painstaking development of laboratory technique. Lucretius did not know them, and they were inaccessible even to Newton and to Dalton. They are a very limited selection from among the phenomena of nature, but not for that the less important. The atom-model which is devised to explain them is at best a partial atom-model; thus far it serves for no other phenomena than these, but these it does interpret with an elegance and a competence quite without precedent among atom-models. I have said that some of these phenomena are explained by conceiving an atom made of a positively-charged nucleus and a family of electrons around it; but this conception is not tenable if unmodified. It can be modified so as to interpret the rest of these phenomena; but this means little by itself. The important fact is this, that the modifications which are demanded appear in some cases to be endowed with a beauty and a simplicity, which indicate that they are the expressions of an underlying principle of Nature. To these the following article will be devoted.

Transatlantic Radio Telephone Transmission[1]

By LLOYD ESPENSCHIED, C. N. ANDERSON
and AUSTIN BAILEY

SYNOPSIS: This paper gives analyses of observations of long-wave transmission across the Atlantic over a period of about two years. The principal conclusions which the data seem to justify are as follows.

1. Solar radiation is shown to be the controlling factor in determining the diurnal and seasonal variations in signal field. Transmission from east to west and west to east exhibit similar characteristics.

2. Transmission in the region bordering on the division between the illuminated and the darkened hemispheres is characterized by increased attenuation. This manifests itself in the sunset and sunrise dips, the decrease in the persistence of high night-time values in summer and the decrease in daylight values during the winter.

3. Definite correlation has been found between abnormal radio transmission and disturbances in the earth's magnetic field. The effect is to decrease greatly the night-time field strength and to increase slightly the daylight values.

4. The limit of the high-night-time value of signal field strength for transatlantic distance is essentially that given by the Inverse Distance Law. The normal daylight field strengths obtained in these tests can be aproximated by a formula of the same form as those earlier proposed but with somewhat different constants.

5. The major source of long wave static, as received in both England and the United States, is indicated to be of tropical origin.

6. In general, the static noise is lower at the higher frequencies. At night the decrease with increase in frequency is exponential. In day-time the decrease with increase in frequency is linear in the range of 15 to 40 kilocycles. The difference between day and night static is, therefore, apparently due largely to daylight attenuation.

7. The effect of the static noise in interfering with signal transmission, as shown by the diurnal variations in the signal-to-noise ratio, is found to be generally similar on both sides of the Atlantic.

8. Experiments in both the United States and England with directional receiving antennas of the wave antenna type show an average improvement in the signal-to-static ratio of about 5 as compared with loop reception.

IT will be recalled that something over two years ago, experiments in one-way radio telephone transmission were conducted from the United States to England.[2] In respect to the clarity and uniformity of the reception obtained in Europe, the results represented a distinct advance in the art over the transatlantic tests of 1915. However, they were carried out during the winter, which is most favorable to radio transmission, and it was realized that an extensive favorable to radio transmission, and it was realized that extensive study of the transmission obtainable during less favorable times would be required before the development of a transatlantic radio telephone service could be undertaken upon a sound engineering basis.

[1] Presented before the Institute of Radio Engineers, May 6, 1925.

[2] "Transatlantic Radio Telephony," Arnold and Espenschied, *Journal of A.I.E.E.*, August, 1923. See also, "Power Amplifiers in Transatlantic Telephony," Oswald and Schelleng, presented before the Institute of Radio Engineers, May 7, 1924.

Consequently, an extended program of measurements was initiated to disclose the transmission conditions obtaining throughout the twenty-four hours of the day and the various seasons of the year. The methods used in conducting these measurements and the results obtained during the first few months of them have already been described in the paper previously mentioned. The results there reported upon were limited to one-way transmission from the United States to England upon the telephone channel. Since then the

Fig. 1

measurements have been extended to include transmission on several frequencies in each direction from radio telegraph stations in addition to the 57 kilocycles employed by the telephone channel.

The present paper is, therefore, in the nature of a report upon the results thus far obtained in work currently under way. It seems desirable to make public these results because of the large amount of valuable data which they have already yielded, and because of the timely interest which attaches to information bearing upon the fundamentals of radio transmission. The carrying on of this extensive measurement program has been made possible through the cooperation of engineers of the following organizations: in the United States—The American Telephone and Telegraph Company and the Bell Telephone Laboratories, Inc., with the Radio Corporation of America and its Associated Companies; in England—The International Western Electric Company, Inc., and the British Post Office.

Measurement Program

The scene of these transatlantic experiments is shown in Fig. 1 The British terminal stations will be seen to lie in the vicinity of London and the American stations in the northeastern part of the United States. The United States transmitting stations are the radio telephone transmitter at Rocky Point, and the normal radio

Fig. 2—Exterior of Riverhead Radio Receiving Station

telegraph transmitters at Rocky Point and Marion, Mass. The measurements of these stations were made at New Southgate and at Chedzoy, England. The British transmitting stations utilized in measuring the east to west transmission were the British Post Office telegraph stations at Leafield and at Northolt. The receiving measurements in the United States were initiated at Green Harbor, Mass., and continued at Belfast, Maine and Riverhead, L. I.

The Riverhead receiving station, shown in Figs. 2 and 3, is typical of the receiving stations involved in the measurement program. The interior view of Fig. 3 shows the group of receiving measurement apparatus at the right and the loop at the left. The three bays of apparatus shown are as follows: That at the left is the receiving set proper which is, in reality, two receiving sets in one, arranged so that one may be set for measurements on one frequency band and the other set upon another band. The set is provided with variable filters which accounts for the considerable number of condenser dials. The second bay from the left contains voice-frequency output apparatus, cathode ray oscillograph and frequency meter. The third bay

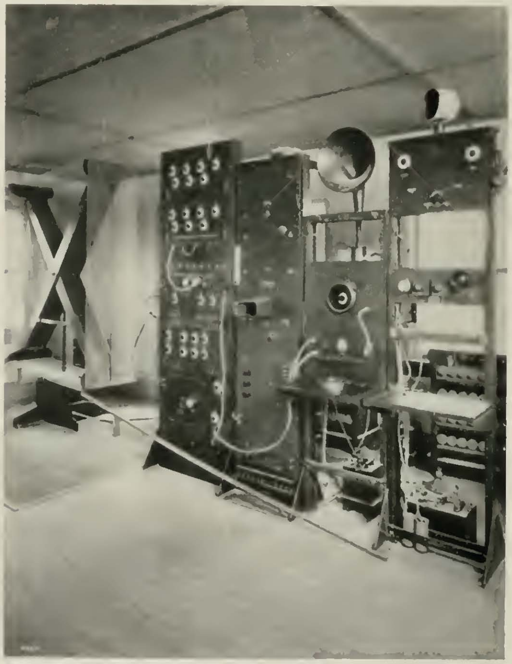


Fig. 3—Interior Riverhead Station

carries the source of local signal and means for attenuating it, and the fourth bay contains means for monitoring the transmission from the nearby Rocky Point radio telephone transmitter.

The measurements are of two quantities: (1), the strength of received field, and (2) the strength of received noise caused by static. The particular frequencies upon which the measurements were taken (given in the chart of Fig. 4) lie in a range between 15 and 60 kc. The arrows indicate the single frequency transmissions which were employed for signal field strength measurements, those at the left indicating the frequencies received in the United States from England, and those at the right, the frequencies received in England from the United States. The black squares in the chart denote the bands in which the noise measurements were taken. In general the measure-

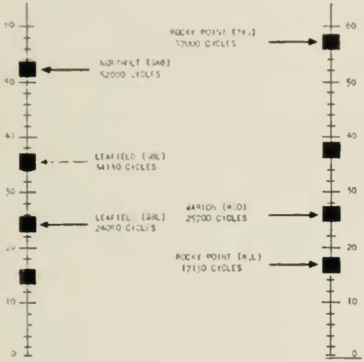


Fig. 4—Frequency distribution of measurements. Black squares denote band in which noise measurement was taken

ments of both field strength and noise have been carried out on both sides of the Atlantic at hourly intervals for one day of each week. The data presented herewith are assembled from some 10,000 individual measurements taken during the past two years in the frequency range noted above. The transmitting antenna current has been obtained for each individual field strength measurement and all values corrected to a definite reference antenna current for each station measured. The data have been subject to careful analysis in order to disclose what physical factors, such as sunlight and the earth's magnetic field, affect radio transmission.

Measurement Methods

Although it will not be necessary to describe in any detail the type of apparatus employed in making these measurements, as this information has already been published,[3] a brief review of the methods involved will facilitate an understanding of the data.

[3] Radio Transmission Measurements, Bown, Englund, and Friis. Proceedings I.R.E., April, 1923.

In general the method employed in measuring the signal field strength is a comparison one. A reference radio-frequency voltage of known value is introduced in the loop antenna and adjusted to give the same receiver output as that from the distant signal. This is determined either by aural or visual means. Under such conditions equal voltages are introduced in the antenna from local and distant sources, and by calculating the effective height of the loop the field strength of the received signal is determined.

In the noise measurements, static noise is admitted through a definite frequency band approximately 2,700 cycles wide. A local radio-frequency signal of known and adjustable voltage is then in-

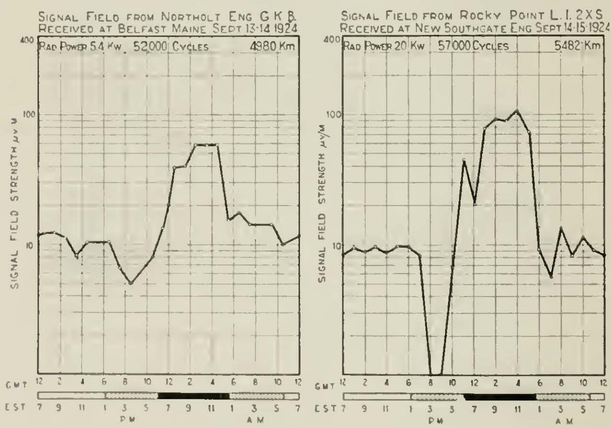


Fig. 5—Diurnal variation in signal field

troduced. The radio-frequency source of this signal is subjected to a continual frequency fluctuation so that the detected note has a warbling sound. This is done in order that the effect of static upon speech can be more closely simulated than by using a steady tone. The intensity of the signal is then adjusted to such a value that further decrease results in a rapid extinction. The comparison signal is then expressed in terms of an equivalent radio field strength. Thus the static noise is measured in terms of a definite reference signal with which it interferes and is expressed in microvolts per meter.

Signal Field Strength

The curves of Fig. 5 are given as examples of the field strength measurements covering a single day's run. The curves have been constructed by connecting with straight lines the datum points of measurements taken at hourly intervals. It will be evident that

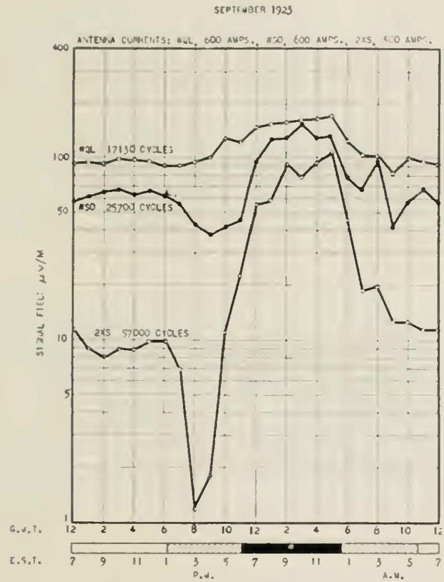


Fig. 6—Monthly average of diurnal variation in signal field transmission from American stations on various frequencies received at New Southgate, England, September, 1923

they portray the major fluctuations occurring throughout the day, but that they are not sufficiently continuous to disclose, in detail, the intermediate fluctuations to which the transmission is subject.

Diurnal Variation. The left-hand curve is for transmission from England to America on 52 kilocycles, and the right-hand one for transmission from America to England on 57 kilocycles. These curves illustrate the fact, which further data substantiate, that both transmissions are subject to substantially the same diurnal variation. The condition of the transatlantic transmission path with respect to daylight and darkness is indicated by the bands beneath the curves. The black portion indicates the time during which the transatlantic path is entirely in darkness, the shaded portions the time during which it is only partially in darkness, and the unshaded portions the time during which daylight pervades the entire path.

The diurnal variation may be traced through as follows:

1. Relatively constant field strength prevails during the daylight period.

2. A decided drop in transmission accompanies the occurrence of sunset in the transmission path between the two terminals.

3. The advent of night-time conditions causes a rapid rise in field strength to high values which are maintained until daylight approaches.

4. The encroachment of daylight upon the eastern terminal causes a rapid drop in signal strength. This drop sometimes extends into a morning dip similar to, but smaller than, the evening dip. After this, relatively steady daylight field strengths again obtain.

Three or four curves similar to Fig. 5 are obtained each month. By taking the average of such curves for the month of September, 1923, the lower curve on Fig. 6 is obtained. The upper curves are for similar averages of measurements made on the lower frequencies. These curves show clearly that the range of the diurnal fluctuation is less for the lower frequencies. This is because of the lesser daylight absorption.

The mechanism by which the transatlantic transmission path is subjected to these daily and seasonal controls on the part of the sun, would be more evident were we enabled to observe the earth from a fixed point in space. We should then be able to see the North Atlantic area plunged alternately into daylight and darkness as the earth rotates upon its axis, and to visualize the seasonal variation of this exposure to sunlight as the earth revolves about the sun. Photographs of a model of the earth showing these conditions have been made, and are shown in Fig. 7. The first condition is that for January, in which the entire path is in daylight. The curve of diurnal variation is shown in the picture and that part which corresponds to the daylight condition is indicated by the arrow. In the next position the earth has rotated so that the London terminal is in darkness while the United States terminal is still in daylight. This corresponds to the evening dip, the period of poorest transmission. With the further rotation of the earth into full night-time conditions for the entire path, the received signal rises to the high night-time values. These high values continue until the path approaches the daylight hemisphere as indicated in the fourth position. As the path enters into sunlight, the signal strength drops with a small dip occurring when sunrise intervenes between the two terminals.

Seasonal Variation. By assembling the monthly average curves for all months of the year, the effect of the seasonal variation on the

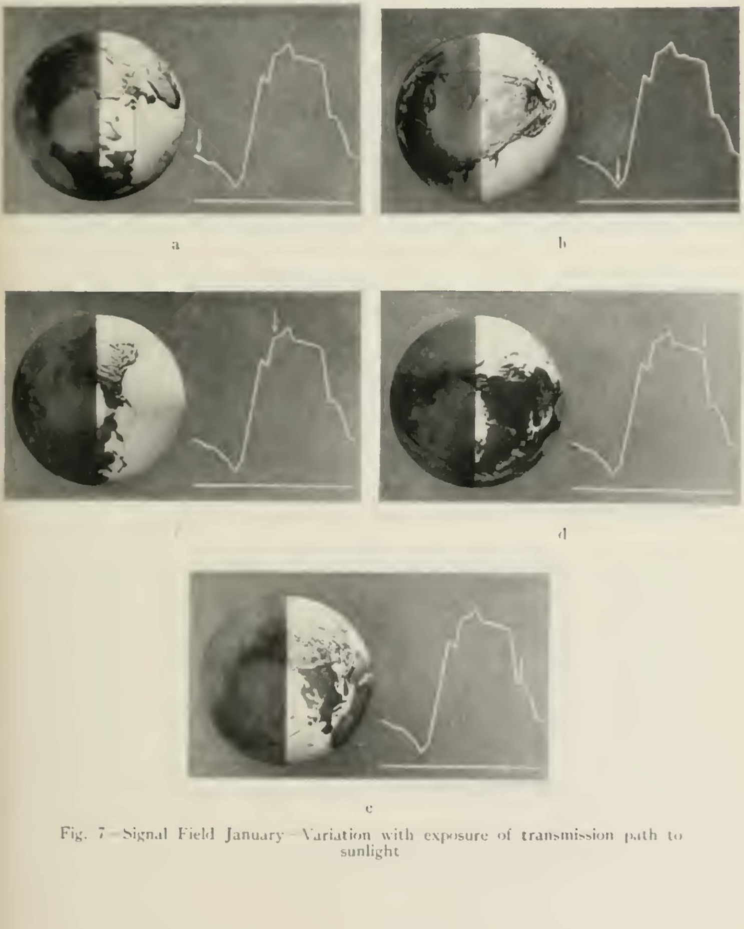


Fig. 7—Signal Field January—Variation with exposure of transmission path to sunlight

diurnal characteristic becomes evident. This is shown in Fig. 8, the data for which actually cover two years.

The outstanding points to be observed in this figure are:

1. The continuance of the high night-time values throughout the year.

2. The persistence of the high night-time values for a longer period in the winter than in the summer months.

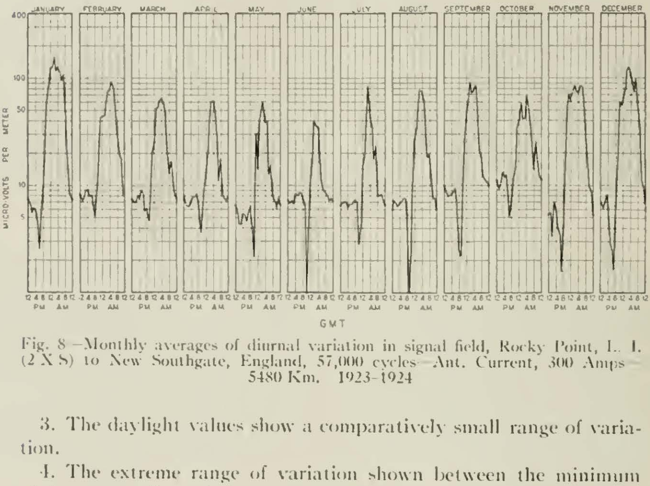


Fig. 8—Monthly averages of diurnal variation in signal field, Rocky Point, L. I. (2 X S) to New Southgate, England, 57,000 cycles—Ant. Current, 300 Amps—5480 Km. 1923–1924

3. The daylight values show a comparatively small range of variation.

4. The extreme range of variation shown between the minimum of the sunset dip and the maximum of the high night-time values is of the order of 1 to 100 in field strength. This is equivalent to 1 to 10,000 in power ratio.

It will be recalled that the cause of the seasonal changes upon the earth's surface resides in the fact that the earth's axis is inclined and not perpendicular to the plane of its orbit about the sun. As the earth revolves about the sun, the sunlit hemisphere gradually extends farther and farther northward in the spring months and by the summer solstice reaches well beyond the north pole, as indicated in Fig. 9. As the earth continues to revolve about the sun, the sunlit hemisphere recedes southward until at the winter solstice it falls considerably short of the north pole and extends correspondingly beyond the south pole. Since the transatlantic path lies fairly high in the northern latitude, it is not surprising that the transmission conditions disclose a decided seasonal influence. The effect of this seasonal influence in shifting the diurnal transmission characteristic is better shown in Fig. 10. This figure consists of the same monthly average diurnal curves as are assembled in Fig. 8, arranged one above the other instead of side by side.

Fig. 9—Signal Field June—Night conditions showing proximity of transmission path to sunlit hemisphere

In particular, there should be noted:

1. The time at which the sunset dip occurs changes with the change in time of sunset.

2. Similarly, the time at which the morning drop in field strength occurs changes with the time of sunrise.

3. The period of high night-time values, bounded between the time of sunset in the United States and the time of sunrise in England, is much longer in the winter than in the summer months.

It is also to be observed that, as a rule, full night-time values of signal field strength are not attained until some time after sunset at the western terminal and that they begin to decrease before sunrise at the eastern terminal. In other words, the daylight effects appear to extend into the period in which the transmission path along the earth's surface is unexposed to direct rays of the sun. The effect of this is that with the advance of the season from winter to summer the time at which the high night-time value is fully attained occurs later and later whereas the time at which it begins to fall off occurs earlier and earlier, until the latter part of April when these two times coincide. At this time, then, the transmission path no sooner comes into the full night-time conditions than it again emerges. As the season further advances into summer, the day conditions begin to set in while the night-time field strength is still rising. The proximity to the daylight hemisphere, which the transatlantic path reaches at night during this season of the year is illustrated in Fig. 9.

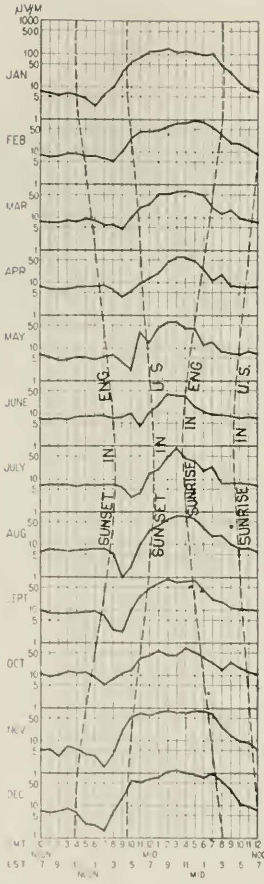


Fig. 10—Monthly averages of diurnal variation of signal field, Rocky Point, L. I. (2 X S) to New Southgate, England; 20.8 K.W. radiated power, 57,000 cycles, 1923–1924

As the sunlit hemisphere recedes southward after the summer solstice a time is reached, about the middle of August, at which the full night-time values are again realized. Beyond this time they are sustained for increasing periods of time. It is of interest to note that at these two times of the year, the last of April and the middle of August, direct sunlight exists over the darkened hemisphere some 500 kilometers above the great circle path.

For all of the conditions noted above, namely, sunset, sunrise, and summer approach of the transmission path to the northern boundary of the night hemisphere, the path lies in a region wherein the radiation

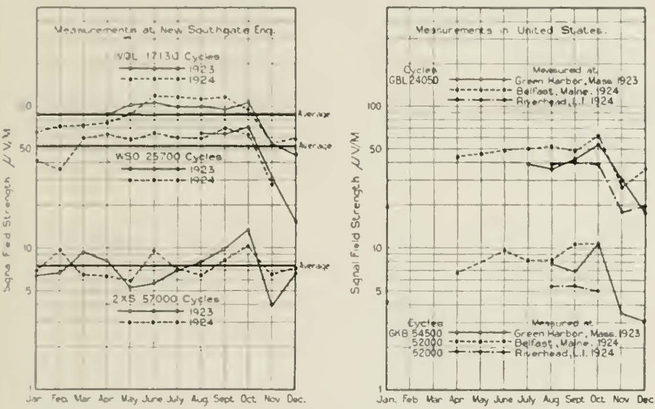


Fig. 11—Monthly averages of daylight field strength

from the sun grazes the earth's surface at the edge of the sun-lit hemisphere. The transmission path also approaches this region during daylight in the winter months, as will be seen by reference to the first position of Fig. 7 for the month of January. The results of measurements for the months of November, December and January for all of the frequencies measured show definite reductions in the daylight field strengths. This reduction is evident in Fig. 8 for the 57-kilocycle transmission, but shows up more strikingly in the curves of Fig. 11. The effect of each of these conditions, in which the transmission path approaches the region in which the solar emanation is tangential to the earth's surface, will be observed to be that of an *increase* in the transmission loss. The fact that in one instance this occurs in daylight would seem to suggest for its explanation the presence of some factor in addition to sunlight, such as electron emission.

Field Strength Formulae. The two major phases of the diurnal variation of signal field strength which lend themselves to possible predetermination are the daylight values and the established night-time values. As to the night-time values our data show, within the limits of experimental error, that the maximum values do not exceed that defined by the inverse distance law. This fact seems to support the viewpoint [4] that the high night-time values are merely the result of a reduction of the absorption experienced during the day. Fig. 11 presents the monthly averages of the *daylight* field strengths for the various frequencies on which measurements were taken. The chart at the left is for reception in England and that at the right for reception in the United States.

The difficulty in predicting by transmission formulae, values to be expected at any one time will be evident and the best that can be expected is to approximate the average. The formulae of Sommerfield, Austin-Cohen and Fuller take the form

$$E\mu v/M = \frac{377HI}{\lambda D} e^{-\frac{\alpha D}{\lambda^x}}$$

where the coefficient $\frac{377HI}{\lambda D}$ represents the simple Hertzian radiation field and the exponential $e^{-\frac{\alpha D}{\lambda^x}}$ the attenuation factor. From theoretical considerations, Sommerfeld (1909) gave $\alpha = .0019$ and $x = 1/3$. In the Austin-Cohen formula α is given as .0015 and $x = 1/2$. Fuller gives $\alpha = .0045$ and $x = 1.4$. The Austin-Cohen formula was tested out experimentally chiefly with data obtained from the Brant Rock station (1911) and from the Arlington station by the U.S.S. *Salem* in February and March, 1913. Fuller derived his .0045 value of α from 25 selected observations from tests between San Francisco and Honolulu in 1911.

An attempt has been made to determine the constants of a formula of the above form which would approximate averages of some 5,000 observed values of field strength over this particular New York to London path and over the frequency range of 17 kc. to 60 kc. For each transmitting station a series of comparatively local measurements were taken to determine the power radiated. By combining these local measurements with the values obtained on the other side

[4] See also "Radio Extension of Telephone System to Ships at Sea," Nichols and Espenschied, Proc. I. R. E., June, 1923, pages 226–227.

of the Atlantic we found that approximately $\alpha = .005$ and $x = 1.25$. The transmission formula then becomes

$$E\mu v / M = \frac{377 HI}{\lambda D} \epsilon^{-\frac{.005D}{\lambda^{1.25}}}$$

or in terms of power radiated

$$E = \sqrt{P}\,\frac{298 \times 10^3}{D} \epsilon^{-\frac{.005D}{\lambda^{1.25}}}$$

where

E = Field strength in microvolts per meter
P = Radiated power in kw.
D = Distance in km.
λ = Wave length in km.

The table shown on next page summarizes the data relative to daylight transmission.

Correlation Between Radio Transmission and Earth's Magnetic Field

In analyzing the measurements we were impressed by the occasional occurrence of marked deviations from the apparent normal diurnal characteristic. A series of measurements which includes an example of this condition is represented in the upper curves of Fig. 12. The curves of the first four days exhibit the normal diurnal characteristic as did the curves of the preceding measurements. The next test of February 25–26 exhibits a marked contrast with that of two days previous. Such abnormality continues in greater or less degree until partial recovery in the test of April 29–30.

Comparison of these data with that of the earth's magnetic field for corresponding days shows a rather consistent correlation. This will be evident from inspection of the magnetic data plotted below in the same figure. Both the horizontal and vertical components of the earth's field are shown. The first decided abnormality occurs February 25–26. The three succeeding periods show a tendency to recover followed by a second abnormality on March 25–26 and again one on April 22–23. It is of interest to note that within limitations of the intervals at which measurements were taken, these periods correspond roughly to the 27-day period of the sun. Coincidences similar to those described above have been found for other periods. Except for this coincidence of abnormal variations in earth's magnetic field and radio transmission, exact correlation of the fluctuations has not been found possible.

TRANSATLANTIC RADIO TELEPHONE MEASUREMENTS

Transmitting Terminal	Receiving Terminal	Freq.	Distance Km.	Power* Radiated Kw.	Daylight Field Strengths Observed			Daylight Field Strengths Calculated		
					1923	1924	Av.	Austin-Cohen	Fuller	This Paper
2 X S	New Southgate, Eng.	57,000	5,482	20.6	7.5 (Aug.-Dec.)	7.65 (Jan.-Nov.)	7.6	6.9	21.2	7.8
WSO	New Southgate, Eng.	25,700	5,282	8.95	48.7 (Apr.-Dec.)	54.6	52.7	16.6	78.5	50.2
WQL	New Southgate, Eng.	17,130	5,482	12	86 (July-Jan.)	87.3	86.8	27.7	116.	86.
GBL	Green Harbor, Mass.	24,050	5,149	4.06	34.2			13.2	59.	39.
	Belfast, Maine	24,050	4,885	4.06		(Apr.-Dec.) 51(?) (Aug.-Dec.)		15.6	54.7	41.8
	Riverhead, L. I.	24,050	5,363	4.06		31.5		11.4	55.2	34.5
GBL	Green Harbor, Mass.	34,130	5,149	4.85	(July-Jan.) 16.1 (Aug.-Dec.)			9.5	41.2	22.6
GKB	Green Harbor, Mass.	54,500	5,241	7.9	6.1			5.6	18.6	7.1
	Belfast, Maine	52,000	4,980	5.4		(Apr.-Oct.) 9.1 (Aug.-Oct.)		6.15	20.	9.05
	Riverhead, L. I.	52,000	5,457	5.4		5.3		4.2	15.	5.9

* Computed from local observations using formula of this paper.

NOTE: Measurements of transmission from Rocky Point (2 X S) on 57,000 cycles measured at Mexico City, July, 1924, give an average daylight field strength of 39.4 mv/M. Calculated value 42.5 mv/M.

The magnetic data have been supplied through the courtesy of the United States Geodetic Survey. Similar data taken in England were obtained from the Kew observatory and show similar results.

The contrast in the diurnal variations of radio transmission before and after the time a magnetic storm is known to have started, is

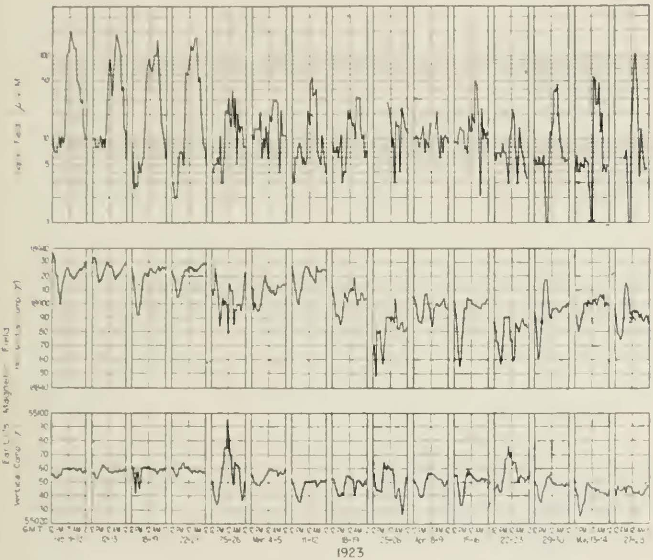


Fig. 12—Correlation of radio transmission and earth's magnetic field—Transmission from Rocky Point, U. S. A. (57,000 cycles) to London, Eng.—Earth's magnetic field measured at Cheltenham, Md., U. S. A.

further brought out in Fig. 13. The lower left-hand curve in this figure superimposes curves of February 22–23 and February 25–26 of the previous figure. Additional cases where such marked changes occur are also shown. It will be seen that similar effects exist on the lower frequency of 17 kc. All of these examples are for days of other than maximum magnetic disturbance. In general the effect is to reduce greatly the night-time values and slightly increase the daylight values. The higher peaks in the daylight field strength of Fig. 11 are due to the high daylight values which prevailed at the time of these disturbances.

Noise Strength

Next to field strength the most important factor in determining the communication possibilities of a radio channel is that of the interfering noise. The extent to which noise is subject to diurnal and seasonal variations is therefore of first order of importance.

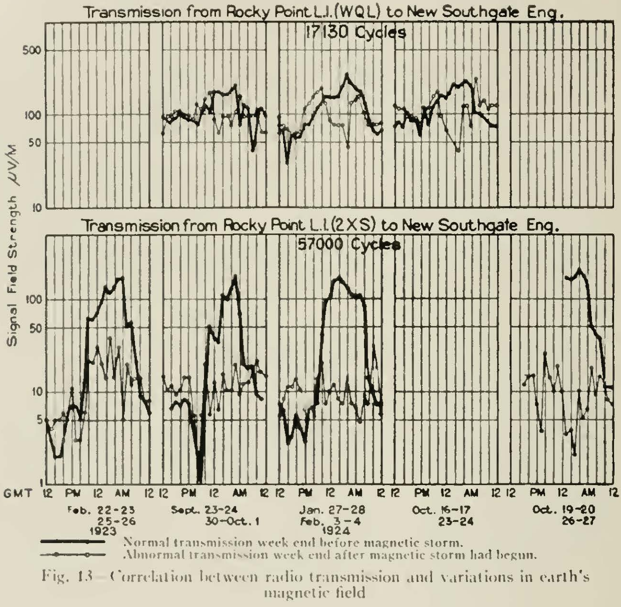


Fig. 13—Correlation between radio transmission and variations in earth's magnetic field

Diurnal Variation. An example of the diurnal characteristic of the noise for both ends of the transatlantic path is given in Fig. 14. One curve is shown for each of the several frequencies measured. The outstanding points to be observed are:

1. The rise of the static noise about the time of sunset at the receiving station, the high values prevailing at night, and the rather sharp decrease accompanying sunrise. The curve for 15 kc. shows the existence of high values also in the afternoon. During the summer months high afternoon values are usual for all frequencies in this

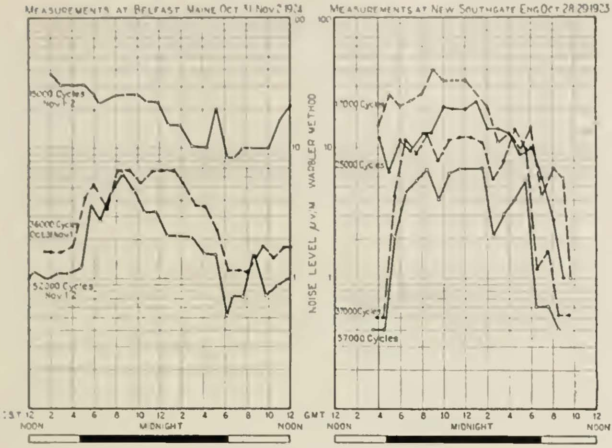


Fig. 14—Diurnal variation in noise

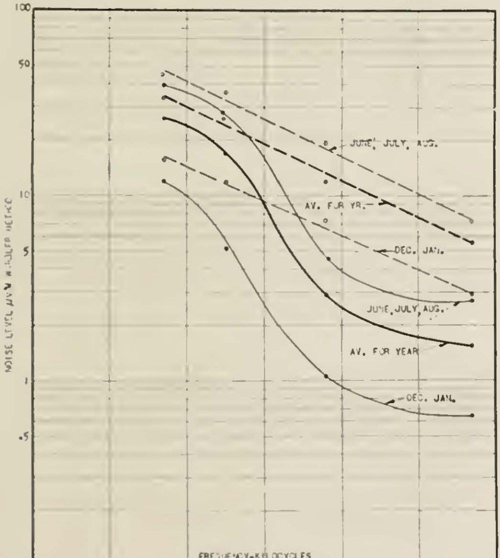


Fig. 15—Frequency distribution of noise, New Southgate, England
Night time———Day time — — —1923–1924

range. They extend later into the fall for the lower frequencies, and hence are in evidence on the date on which these measurements were taken, October-November.

2. In general the noise is greater the lower the frequency.

Noise as a Function of Frequency and of Receiving Location. The distribution of static noise in the frequency range under consideration is depicted in Fig. 15 for the case of reception at New Southgate, England. The set of full-line curves is for daylight reception and the set of dash-line curves for night-time reception. The values obtaining during the transition period between day and night have been excluded. For both conditions three curves are shown, one the average of the summer months, another the average of winter months and the third, the heavy line, the average for the entire year. The curves represent averages for all of the measurements taken during both 1923 and 1924. In considering curves of this type it should be remembered that they represent an average of a wide range of conditions and at any one time the distribution of static may differ widely from that indicated by the curves. Also it should be realized that the extreme difference between winter and summer static is much greater than the difference between the averages.

A similar study of frequency distribution was made at two locations in the United States, Belfast and Riverhead. The results obtained at these two locations together with those for New Southgate, England, are presented in Fig. 16 for a period during which data were obtained for all three places. The similarity of the three sets of curves shows that there is an underlying cause common to both sides of the Atlantic which may account for the difference between the daytime and night-time static on the longer waves. It will be evident from the curves that for frequencies around 20 kc. there is not very much difference between the day and night static noise but that at the higher frequencies in the range studied, the daylight values become considerably less than the night-time values. Actually the divergence between the night-time and the daytime noise curves up to about 40 kc. is an exponential one. This suggests that the lowering of the daylight values may be largely due to the higher absorption which occurs in the transmission medium during the day. There is a further interesting point to be noted concerning both figures, namely, that the night-time values decrease exponentially with increase in frequency. Since these night-time values are but little affected by absorption in the transmitting medium, the distribution of the static energy as received, also roughly represents the distribution of the static power generated.

The curves of Fig. 16 show also the substantial difference in the noise level which exists at the three receiving points. As has been experienced in practice, the New Southgate curve indicates that England is less subject to interference than northeastern United

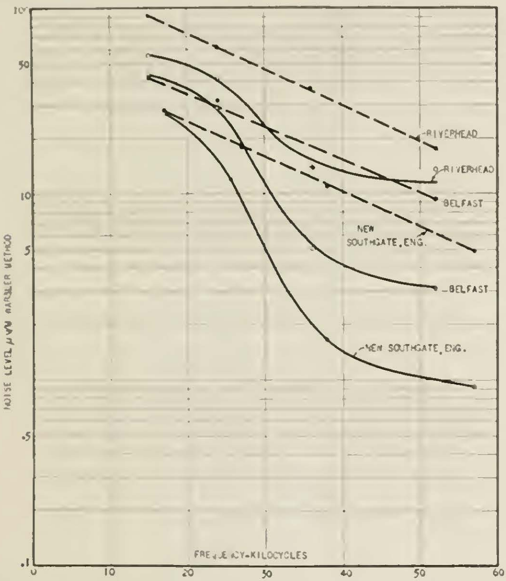


Fig. 16—Frequency distribution of noise, New Southgate, Eng., Belfast, Maine Riverhead, L. I.—Night time ——— Day time — — — Aug.–Dec., 1924

States. In the United States the superiority of Belfast over Riverhead is also consistent with the better receiving results which in general have been experienced in Maine. There should be noted also the fact that the curves for these three locations lie one above the other in the inverse order of the latitudes. This is in keeping with other evidence which points towards the tropical belt as being a general center of static disturbance on the longer wave lengths. Further evidence on this point is presented below in connection with the seasonal variations of noise.

Seasonal Variation. Curves showing the diurnal variation in noise level for each month of the year together with the variation in time of sunset and of sunrise, are shown in Fig. 17. Each curve is the average of all the measurements taken during that particular month in 1923 and 1924. The diurnal variations are generally similar for the different months in respect to the high night-time values which are limited to the period between the times of sunset and sun-

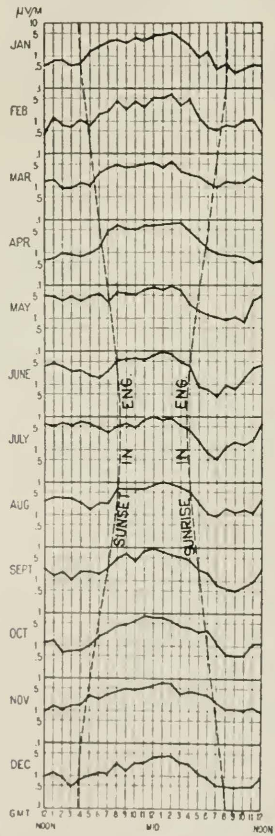


Fig. 17—Monthly averages of diurnal variation of noise, New Southgate, England—57,000 cycles—1923 1924

rise in England. There is a certain deviation, however, which it is well to point out. During the summer months the rise in night-time static starts several hours before and reaches high values at about sunset in England, whereas in the winter-time, the night-timestatic begins to rise at about sunset and reaches high values several hours later. A similar effect is observed for the sunrise condition wherein

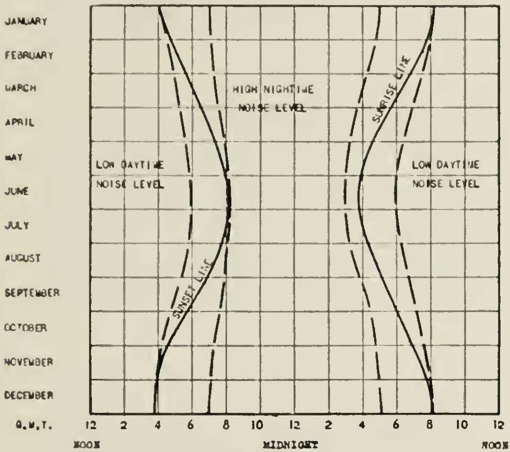
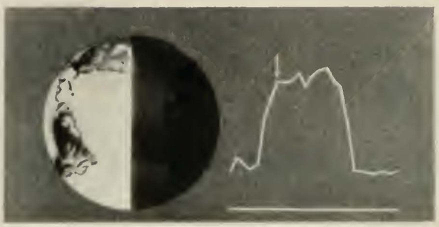


Fig. 18—Seasonal variation in distribution of daytime and night time noise with respect to sunset and sunrise, New Southgate, England—1923–1924

the reduction of static sets in during the summer months about the time of sunrise, reaches low daylight values several hours later, and in the winter the reduction commences several hours before sunrise and reaches low daylight values at sunrise. In other words, the rise to high night-time values occurs earlier with respect to sunset in the summer than in the winter, and conversely the fall from high night-time static to the lower daylight values occurs later with respect to sunrise, in the summer than in the winter.

This is more definitely brought out in Fig. 18 which combines the data for all frequencies measured. The dash-lines associated with the sunset curves, delineate the beginning and the attainment of the night-time increases and those associated with the sunrise curve delineate the beginning and the attainment of the low daylight values. This discloses the fact that sunset and sunrise at the receiving

a



b

c

Fig. 19—Noise at New Southgate, England, in January—Variation with exposure of equatorial belt to sunlight

point does not completely control the rise and fall of the high night-time static. It has been found that the discrepancy can be accounted for if sunrise and sunset are taken with respect to a static transmission path as distinguished from the receiving point alone, and if the assumption is made that the effect of sunlight upon the static transmission path is similar to that on usual radio transmission.

Major Regional Source of Static Noise

A broader conception as to the causes underlying the diurnal and seasonal variation is obtained by considering the time of sunset and sunrise over a considerable area of the earth's surface. Fig. 19 shows a series of day and night conditions for three representative parts of the diurnal noise characteristic at England for January. It will be seen that the rise to high night values does not begin until practically the time of sunset in England with over half of Africa still in daylight. By the time the high night-time values are reached, as indicated in the second phase, darkness has pervaded all of the equatorial belt to the south of England. Incidentally at this time sunset occurs between the United States and England, resulting in very poor signal transmission. The third phase of this series shows the noise having just reached the low daytime value and, although the sun is just rising in England, the African equatorial belt is in sunlight, subjecting the static transmission path to high daylight attenuation.

The sunset conditions which existed for the afternoon and evening of the day upon which the diurnal measurements of Fig. 14 were taken are shown in Fig. 20. The hourly positions of the sunset line are shown in relation to the evening rise of static in London. The coincidence between the arrival of sunset in London and the *start* of the high night-time noise on the higher frequencies is evident. By the time the high night-time values are reached, about 7 o'clock G.M.T., the equatorial belt to the south of London is in darkness.

Fig. 21 shows the sunrise conditions in relation to the decrease in static from the high night-time values to the lower daylight values. The decline starts about 5 or 6 o'clock an hour or two before sunrise, and is not completed until several hours later, at which time daylight has extended over practically the entire tropical belt to the south of England which corresponds in general to equatorial Africa.

Another fact presented in the previous figures which appears to be significant in shedding light upon the source of static, is that noise on the lower frequencies rises earlier in the afternoon and persists later into the morning than does the noise on the higher frequencies. This could be accounted for on the basis that the limits of the area from which the received longer wave static originates, extend farther along the equatorial zone than they do for the higher frequencies.

The inclination of the shadow line on the earth's surface, which is indicated in the previous figure for October 28, shifts to a maximum at the winter solstice, recedes to a vertical position at the equinox and then inclines in the opposite direction. These several positions are illustrated in Fig. 22. The set of three full lines to the right shows the position which the sunset shadow line assumes upon the earth's surface for each of three seasons—winter solstice, equinox, and summer solstice. Likewise, the dash-line curves show the position assumed by the sunrise line for the corresponding seasons. The

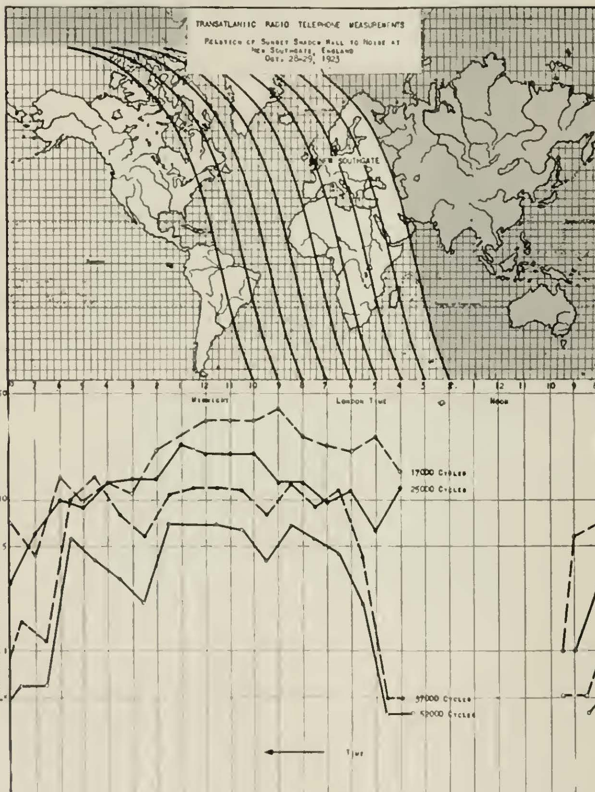


Fig. 20—Relation of sunset shadow wall to noise at New Southgate, England Oct. 28–29, 1923

particular time of day for which each of the sunset curves is taken, is that at which the static in London *begins* to increase to large night values. In winter, this occurs about sunset, at the equinox about one hour earlier, and in summer about two hours earlier, as illustrated in Fig. 18. Correspondingly, the time for which each of the sunrise curves is taken, is that at which the high night-time values have reached the lower daylight values. From Fig. 18 it will be evident

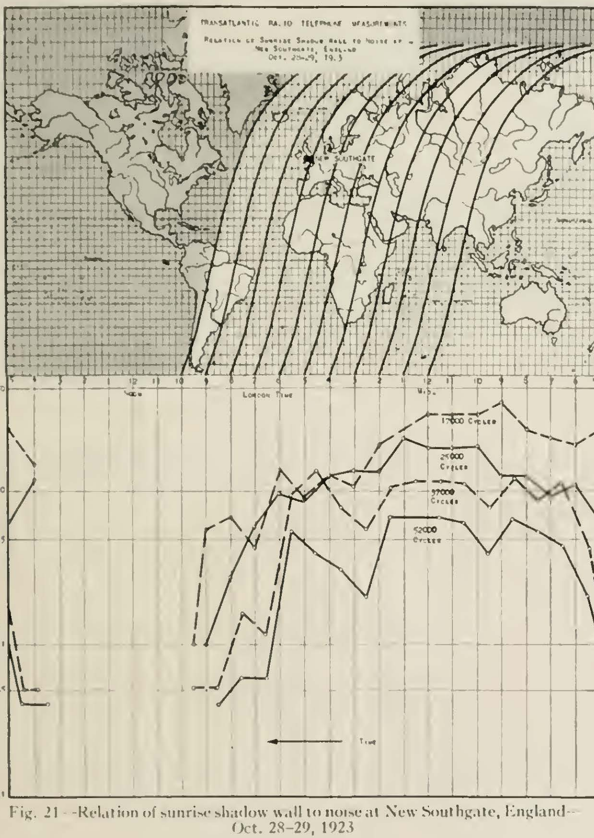


Fig. 21—Relation of sunrise shadow wall to noise at New Southgate, England—Oct. 28–29, 1923

that this occurs during the winter at about sunrise, at the equinox about an hour later, and during the summer some two hours later.

It will be observed that the two sets of curves, one for sunset and the other for sunrise, intersect at approximately the same latitude, the sunset curves southeast and the sunrise curves southwest of

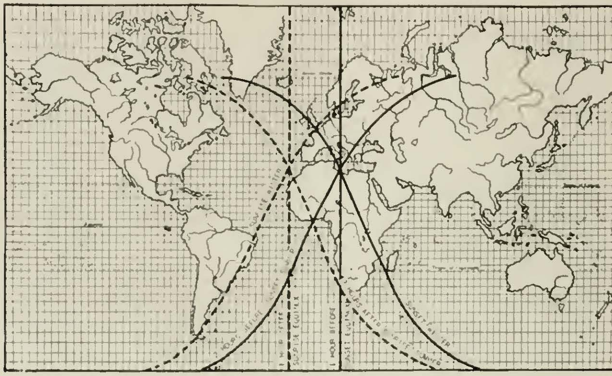


Fig. 22—Position of sunset lines at sunset dip and sunrise lines at sunrise dip in noise level in England for various seasons

England. If it is assumed that the effect of the shadow wall upon the transmission of static is similar to that upon signal transmission across the Atlantic, namely, the high night-time values commence when the shadow wall is approximately half-way between the terminals, the crossing of the lines upon the chart may be taken as having significance in roughly determining the limits of the tropical area from which the major static originates. The crossing of the sunset lines indicates that the eastern limit of the area which contributes most of the static to England is equatorial East Africa. The crossing of the sunrise lines indicates that the corresponding western limit is somewhere in the South Atlantic, between Africa and South America. In other words, from these data the indications are that there is a more or less distinct center of gravity of static, which extend along the tropical belt, and that most of the long-wave static which affects reception in England comes from the equatorial region to the south of England, namely, equatorial Africa. This is exclusive of the high afternoon static prevailing during the summer months.

The data obtained in the United States indicate that generally similar conditions exist there as to the relation between sunset and sunrise path and the major rise and fall of static. This relationship is shown in Fig. 23, which shows in the upper half the course of the

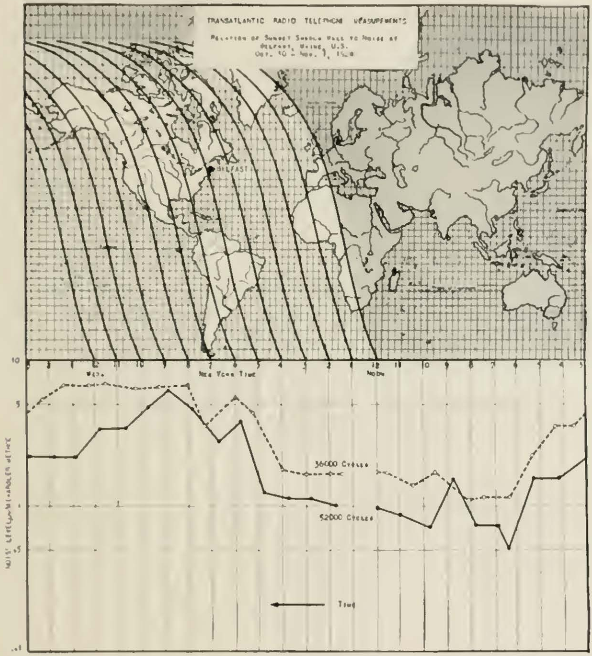


Fig. 23—Relation of sunset shadow wall to noise at Belfast, Maine, U. S.—Oct. 30—Nov. 1, 1924

night-time belt as it proceeds from Europe to America and the corresponding rise in the static noise. The noise level curves are the same as those shown in Fig. 14 for reception at Belfast, Maine. The rise commences about one hour before and continues for one hour or so after sundown. This is for the fall season of the year. A similar chart for the sunrise conditions is given in Fig. 24. Although high night-time values started to fall off some five hours before sundown in Belfast, the more rapid drop was within some two hours in advance. While these curves are for but a single day, they are fairly representative of the average of a greater amount of data. The change in the inclination of the sunset-sunrise curves with the

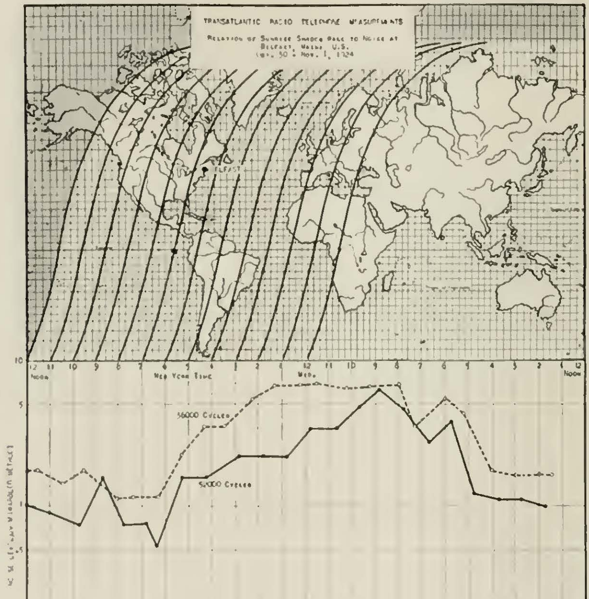


Fig. 24—Relation of sunrise shadow wall to noise at Belfast, Maine, U. S.—Oct. 30–Nov. 1, 1924

season of the year effects changes for American reception somewhat similar to those shown for reception in England, except that for the summer months the coincidence is less definite. It may be that this is because of the somewhat lower latitude of the United States terminal and of the reception of a greater proportion of the static from the North American continent.

In general, therefore, the American results accord with those obtained in England in indicating quite definitely that a large proportion of the static received on the longer waves is of tropical origin.

Signal to Noise Ratio

It is, of course, the ratio of the signal to noise strength which determines the communication merit of a radio transmission channel.

Variation with Frequency. A comparison for representative summer and winter months is given in Fig. 25 of the signal-to-noise ratio

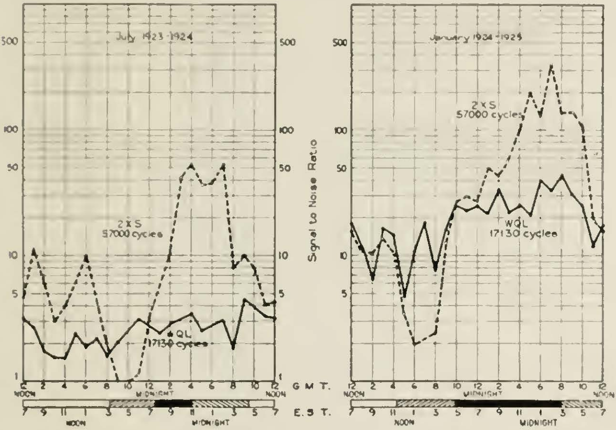


Fig. 25—Variation of signal to noise ratio with frequency. Corrected to same antenna input power (68.5 KW) in Rocky Point antenna—Reception at New Southgate, England

for the two extreme frequencies measured. Both of these transmissions were effected from the same station, Rocky Point, and similar antennae were employed. Comparison is made of the overall transmission by correcting the values of the two curves to the same antenna power input, the power of both channels being scaled down to 68 kilowatts, the power used in the telephone channel during the early parts of the experiment. This chart shows clearly the greater stability in signal to noise ratio obtainable on the lower frequency channel. While for certain periods of the day the higher frequency gives a much better ratio, it is subject to a much more severe sunset

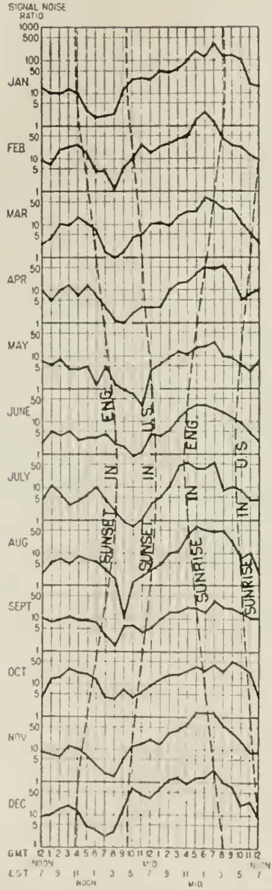


Fig. 26—Monthly averages of diurnal variation of signal to noise ratio; Rocky Point, L. I. (2 X S) received at New Southgate, England; 20.8 KW radiated Power —57,000 cycles—5480 Km—1923-24

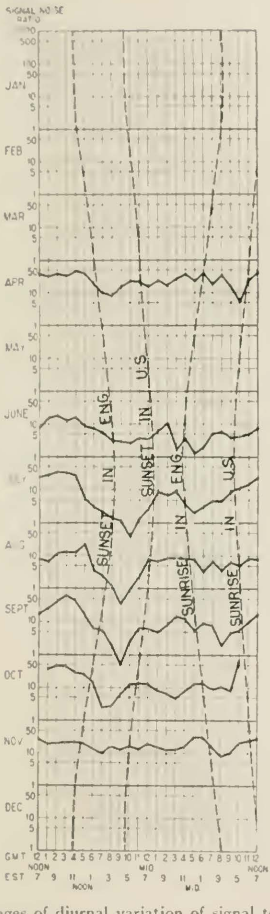


Fig. 27—Monthly averages of diurnal variation of signal to noise ratio, Northolt, Eng. (GKB) received at Belfast, Maine—20.8 KW radiated power—4980 Km—52,000 cycles—1924

decline than is the lower frequency. During the summer time, afternoon reception in England is better on the higher frequency channel. This is because of the considerably greater static experienced at this time on the lower frequency. The higher signal-to-noise ratio prevailing during the winter month of January as compared with the summer month of July is evident. This is due primarily to higher summer static.

Seasonal Variation in England and United States. For the 57-kilocycle channel there is shown in Fig. 26, for each month of the year, signal-to-noise ratios of two years' data. These show a distinct dip corresponding to the sunset dip of the signal field strength. The night-time values are generally high in accordance with the high night-time signal strength but the maximum values are shifted toward the time of sunrise. This is due to the fact that the noise rises earlier in the afternoon and declines earlier in the morning than do the corresponding variations in signal strength.

Fig. 27 presents the signal-to-noise ratios for such data as have thus far been obtained upon transmission from England to the United States on a frequency of 52 kilocycles. The low values obtained about sunset are, of course, due to the evening dip in field strength. In general, the night-time ratios do not reach high values as do those for England because the early morning signal field strength begins to fall off while the noise level is still high. Comparisons of the signal-to-noise ratios obtained at New Southgate and at Belfast show that the Belfast values are somewhat higher for that part of the day, corresponding to forenoon in the United States and afternoon in England. This is because the forenoon static in the United States is lower than the afternoon static in England.

Directive Receiving Antennae

The picture which has been given of the transmission of static northward from the tropical belt suggests that the signal-to-noise ratio might be materially improved by the use of directional receiving systems. This is, of course, what has actually been found to be the case in commercial transatlantic radio telegraphy wherein the Radio Corporation has made such effective use of the wave antenna devised by Beverage. The expectations are confirmed by measurements which have been made in the present experiments using such wave antennae.

A year and a half ago the British Post Office established a wave antenna with which to receive from the Rocky Point radio telephone transmitter. More recently a program of consistent observations in directional reception of east-to-west transmission was also undertaken in which were employed, wave antennae built by the Radio Corporation of America for radio telegraph operation upon lower frequencies.

An indication of the improvement which the wave antenna gives in signal-to-noise ratio is had by reference to Fig. 28. The set of

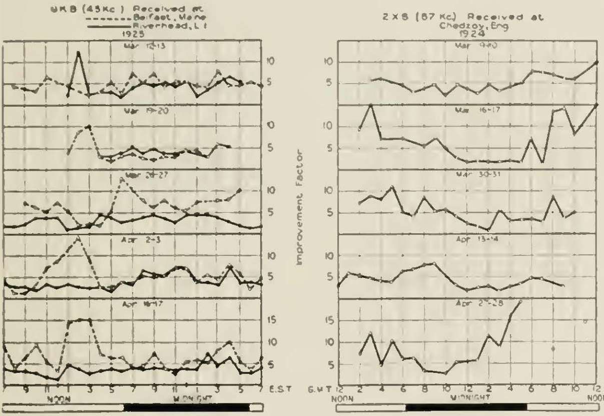


Fig. 28—Improvement in signal noise ratio of wave antenna over loop reception

curves to the right is for reception at Chedzoy, England, and those at the left for reception at Belfast and Riverhead in the United States. The improvement is measured in terms of the signal-to-noise ratio obtained on the wave antenna, divided by the signal-to-noise ratio measured on the loop. For the particular days and frequency indicated, the improvement in England will be seen to vary over a considerable range, averaging about 5. Data for reception in England is for 1924 while that for the United States is for the corresponding period of 1925. The United States results will be seen to be generally similar to those obtained in England. Although these experiments are still in an early stage, the results do give a measure of the order of improvement which can be expected.

Test of Words Understood. Perhaps the most convincing measure of the efficiency of directional receiving systems for transatlantic transmission is the improvement effected in the reception of intelligible words. Fig. 29 shows the improvement which the wave antenna in England has made in the ability to receive certain test words spoken from Rocky Point. For this purpose there was transmitted from Rocky Point a list of disconnected words. A record

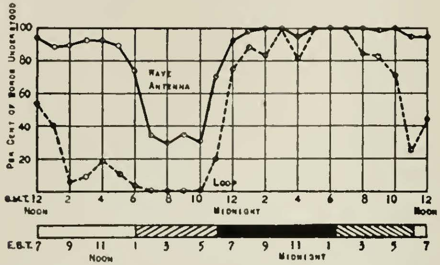


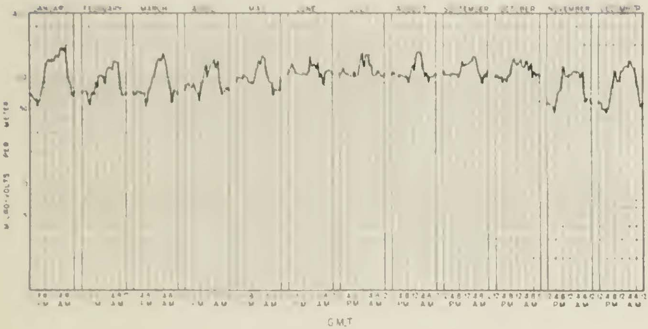
Fig. 29—Comparison of reception on wave antenna and loop. Per cent of words understood—Reception of Rocky Point (2 X S) at Chedzoy, England, March, 1924

was made at Chedzoy of the percentage of the words understood for reception on the loop and on the wave antenna. This constitutes a convenient method of rough telephone testing. It will be appreciated, however, that it would be possible to understand a greater proportion of a conversation than is represented by these results. The curves show that it was possible to receive, for example, 80% of the words for but 9 of the 24 hours on the loop, whereas with the wave antenna reception continued for 18 hours.

APPENDIX

Transatlantic Radio Telephone Measurements
1923, 1924, 1925

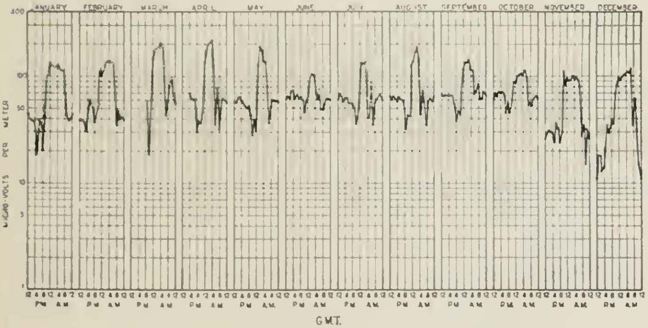
Month by Month Record of Noise and Field Strength



Monthly Averages of Diurnal Variation of Signal Field Strength
Rocky Point, L. I., U. S. A. (WQL) Measured at New Southgate, England
Corrected to 600 Amperes Antenna Current

5,480 Km. 17,130 Cycles

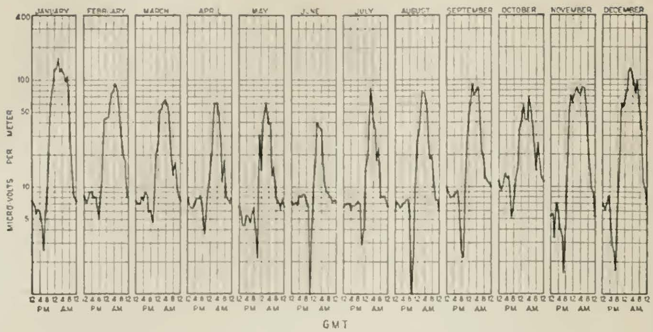
April, 1923—Feb., 1925



Monthly Averages of Diurnal Variation of Signal Field Strength
Marion, Mass., U. S. A. (WSO) Measured at New Southgate, England
Corrected to 600 Amperes Antenna Current

5,280 Km. 25,700 Cycles

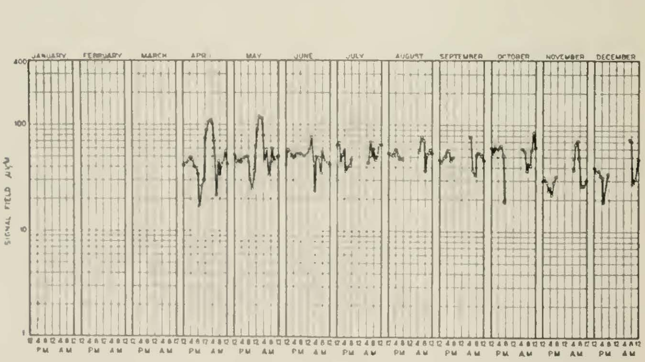
Aug., 1923—Feb., 1925



Monthly Averages of Diurnal Variation of Signal Field Strength
Rocky Point, L. I., U. S. A. (2XS) Measured at New Southgate, England
Corrected to 300 Amperes Antenna Current

5,480 Km. 57,000 Cycles

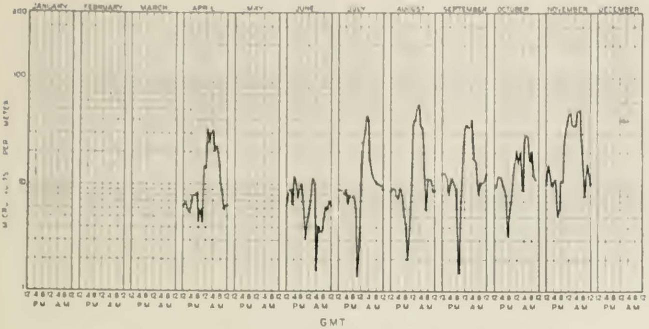
Jan., 1923—Dec., 1924



Monthly Averages Diurnal Variation of Signal Field Strength
Leafield, England (GBL) Measured at Belfast, Maine
Corrected to 300 Amperes Antenna Current

4,980 Km. 24,050 Cycles

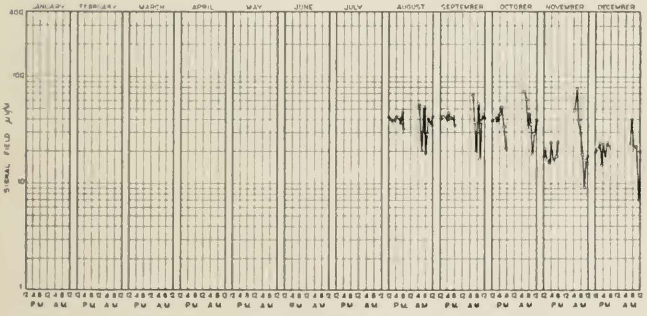
1924



Monthly Averages Diurnal Variation of Signal Field Strength
Northolt, England (GKB) Measured at Belfast, Maine
Corrected to 100 Amperes Antenna Current

4,885 Km. 52,000 Cycles

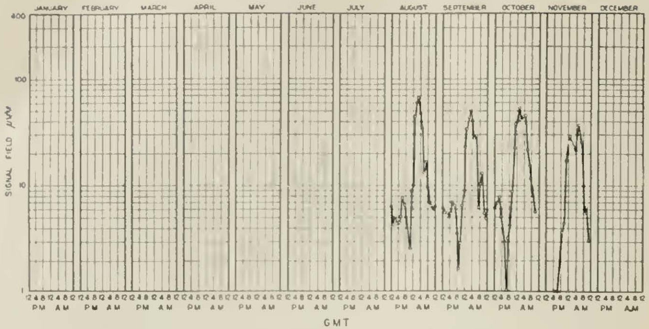
1924



Monthly Averages Diurnal Variation of Signal Field Strength
Leafield, England (GBL) Measured at Riverhead, L. I.
Corrected to 300 Amperes Antenna Current

5,360 Km. 24,050 Cycles

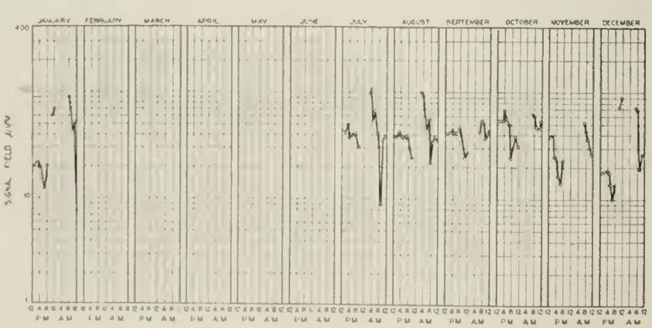
1924



Monthly Averages Diurnal Variation of Signal Field Strength
Northolt, England (GKB) Measured at Riverhead, L. I.

5,460 Km. 52,000 Cycles

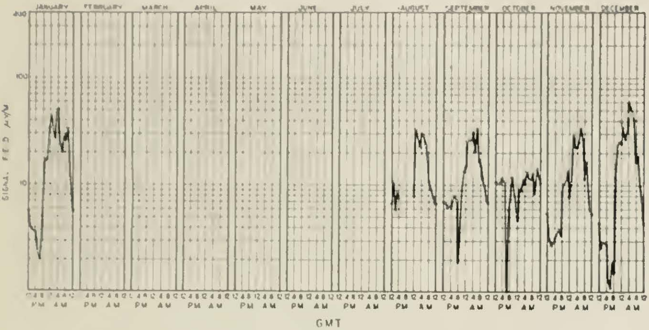
1924



Monthly Average of Diurnal Variation of Signal Field Strength
Leafield, England (GBL) Measured at Green Harbor, Mass.
Corrected to 300 Amperes Antenna Current

5,150 Km. 24,050 Cycles

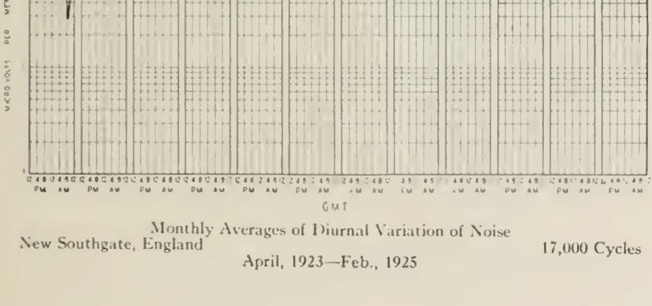
July, 1923—Jan., 1924



Monthly Average of Diurnal Variation of Signal Field Strength
Northolt, England (GKB) Measured at Green Harbor, Mass.
Corrected to 100 Amperes Antenna Current

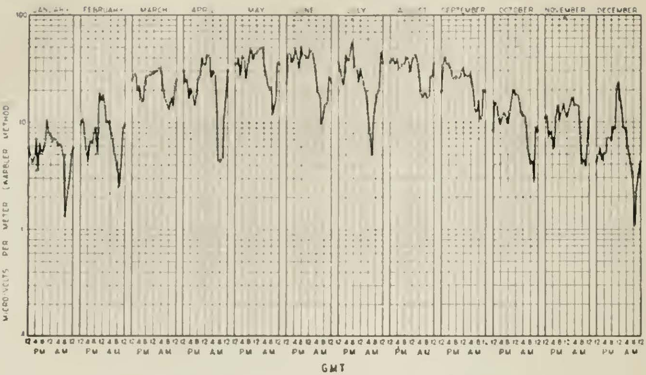
5,240 Km. 54,500 Cycles

Aug., 1923 -Jan., 1924

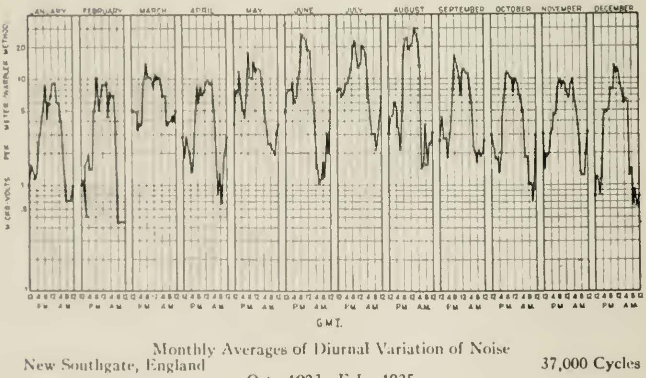


Monthly Averages of Diurnal Variation of Noise
New Southgate, England 17,000 Cycles

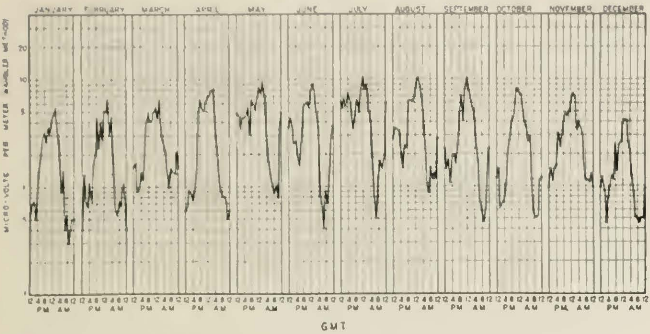
April, 1923—Feb., 1925



Monthly Averages of Diurnal Variation of Noise
New Southgate, England 25,000 Cycles
Aug., 1923—Feb., 1925



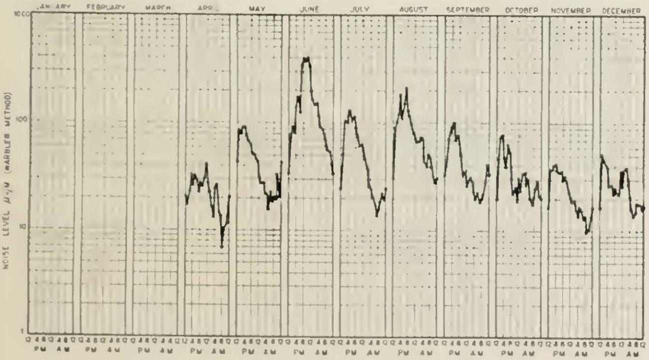
Monthly Averages of Diurnal Variation of Noise
New Southgate, England 37,000 Cycles
Oct., 1923—Feb., 1925



Monthly Averages of Diurnal Variation of Noise

New Southgate, England — 57,000 Cycles

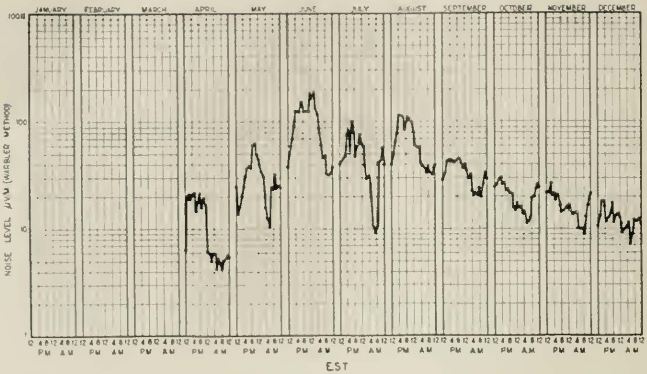
1923–1924



Monthly Average of Diurnal Variation of Noise

Belfast, Maine — 15,000 Cycles

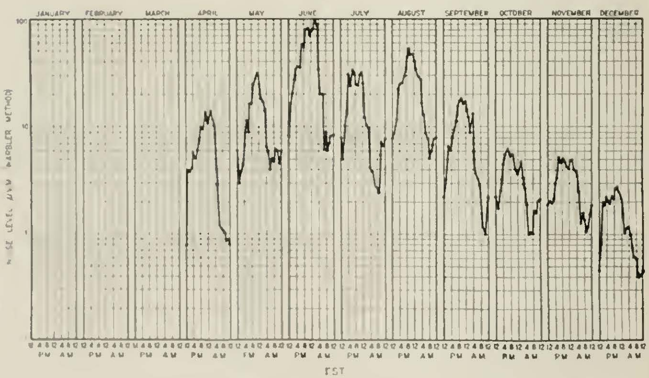
1924



Monthly Average of Diurnal Variation of Noise

Belfast, Maine — 24,000 Cycles

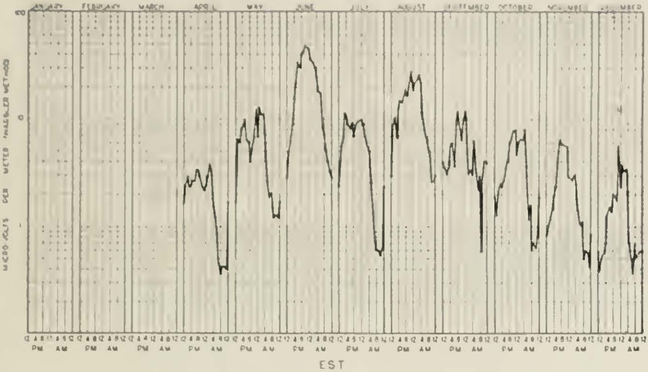
1924



Monthly Average of Diurnal Variation of Noise

Belfast, Maine — 36,000 Cycles

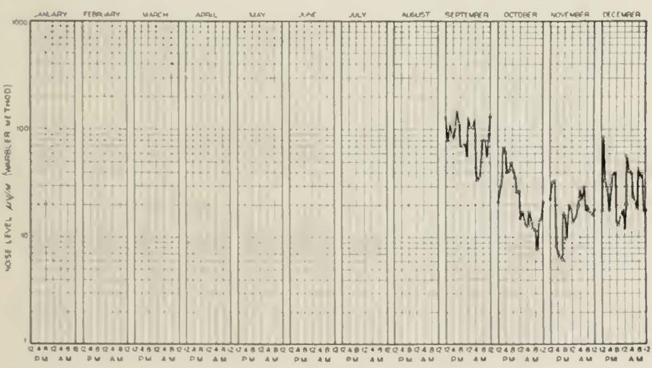
1924



Monthly Average of Diurnal Variation of Noise

Belfast, Maine | 52,000 Cycles

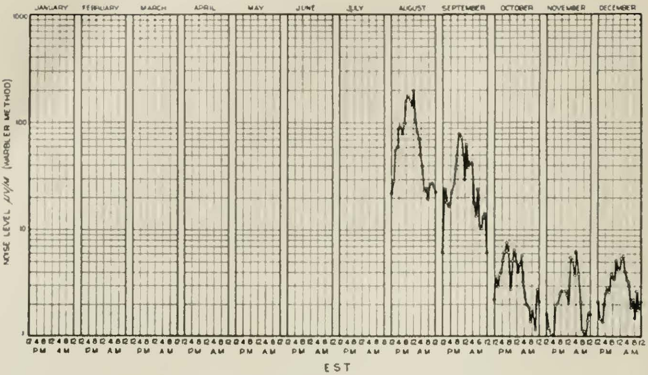
1924



Monthly Average of Diurnal Variation of Noise

Riverhead, L. I. | 15,000 Cycles

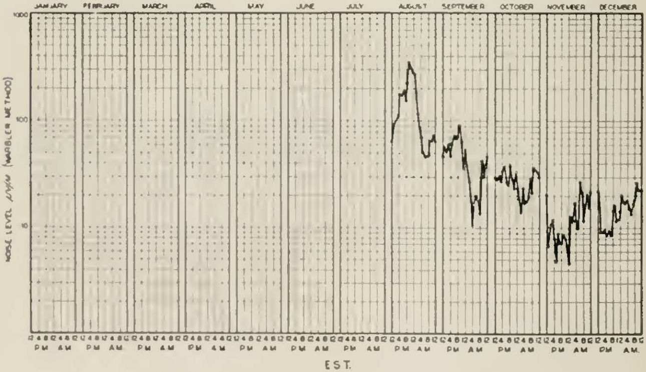
1924



Monthly Average of Diurnal Variation of Noise

Riverhead, L. I. 36,000 Cycles

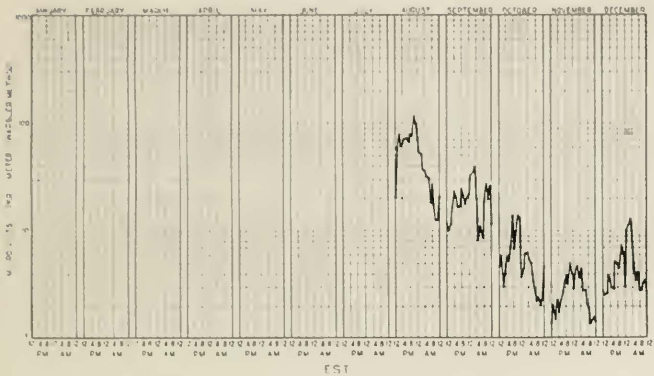
1924



Monthly Average of Diurnal Variation of Noise

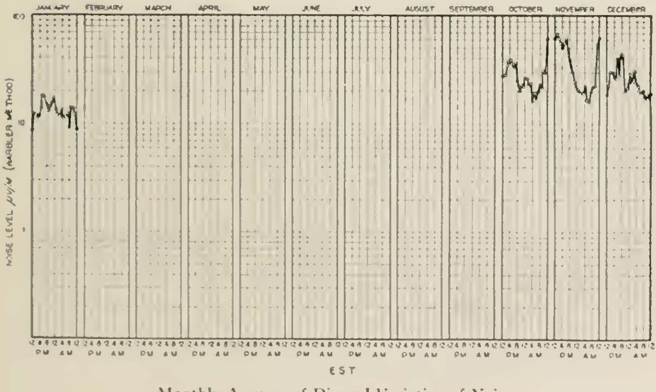
Riverhead, L. I. 24,000 Cycles

1924



Monthly Average of Diurnal Variation of Noise

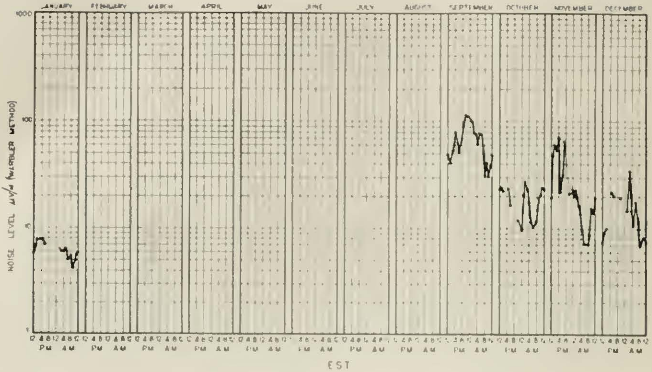
Riverhead, L. I. 52,000 Cycles



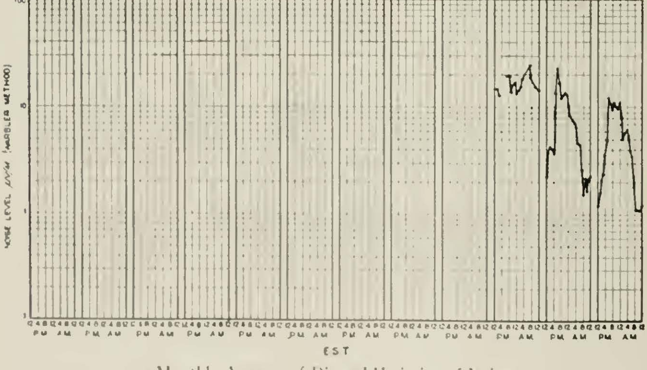
Monthly Average of Diurnal Variation of Noise

Green Harbor, Mass. 15,000 Cycles

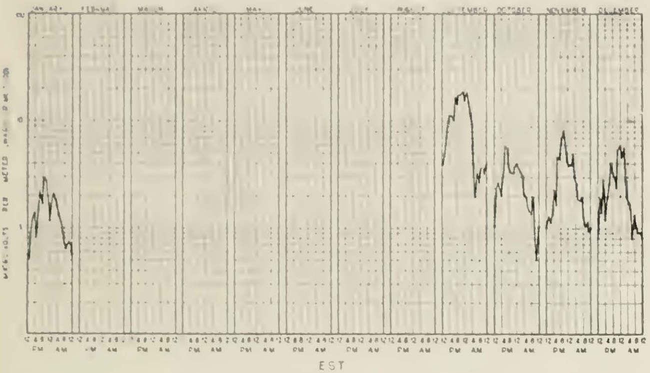
Oct., 1923—Jan., 1924



Monthly Average of Diurnal Variation of Noise
Green Harbor, Mass. 24,000 Cycles
Sept., 1923—Jan., 1924



Monthly Average of Diurnal Variation of Noise
Green Harbor, Mass. 34,000 Cycles
1923



Monthly Average of Diurnal Variation of Noise

Green Harbor, Mass. 55,000 Cycles

Sept., 1923—Jan., 1924

Abstracts of Bell System Technical Papers Not Appearing in this Journal

Radioactivity.[1] A. F. KOVARIK and L. W. MCKEEHAN. This review of progress in radioactivity forms one of a series of monographs prepared by committees of the National Research Council. It outlines the experimental and theoretical advances in the subject since 1916, the date of the last compendium. The section headings are: I. Introduction, II. Radioactive Transformations, III. Alpha-Rays, IV. Beta-Rays, V. Gamma-Rays, VI. Nuclear Structure and Radioactive Processes, VII. Radioactivity in Geology and Cosmology, VIII. The Effects of Radioactive Radiations upon Matter. The references to periodical literature are particularly detailed.

Echo Suppressors for Long Distance Telephone Circuits.[2] A. B. CLARK and R. C. MATHES. This paper gives a brief description of a device which has been developed by the Bell System for suppressing "echo" effects which may be encountered under certain conditions in telephone circuits which are electrically very long. The device has been given the name "echo suppressor" and consists of relays in combination with vacuum tubes which are operated by the voice currents so as to block the echoes without disturbing the main transmission.

A number of echo suppressors have been operated on commercial telephone circuits for a considerable period, so that their practicability has been demonstrated.

The Telephone Transmission Unit.[3] DR. F. B. JEWETT. The adoption by the Bell System of the TU as a telephone transmission unit aroused considerable active discussion in foreign circles, namely, by Colonel Purves, Engineering Chief of the British Post Office Department, and Dr. Breisig of the German Telephone Administration. In this short paper, Dr. Jewett explains certain words and expressions which, when accurately defined, he believes will eliminate misinterpretations such as seem to have led to the controversies over the Bell System TU.

Dr. Jewett also points out that the numerical size for a transmission unit is controlled by two factors, first, the magnitude should be such that computation is convenient, and second, the magnitude should be such as to permit telephone engineers and operating people to most

[1] Bulletin National Research Council, Vol. 10, part 2, March, 1925, 203 pages.

[2] Journal A. I. E. E., Vol. 44, page 618, 1925.

[3] London Electrician, Vol. 94, page 562, 1925

sample itself is balanced by passing a measured current through a third coil. The applied field and the induced magnetization are then proportional to the electric currents passed through the magnetizing coil and the balancing coil, respectively. A hysteresis loop is shown, obtained from an iron wire weighing 3 mg.

An Explanation of Peculiar Reflections Observed on X-Ray Powder Photographs.[8] RICHARD M. BOZORTH. There has been previously reported (J. O. S. A. and R. S. I. 6,989-97; 1922) the existence of "anomalous" reflections of X-rays, observed when analyzing substances by the method of Debye-Scherrer and Hull. These reflections are now explained in accordance with the well-known laws governing X-ray reflections. It is shown that the molybdenum X-ray spectrum as ordinarily used, although it is filtered by zirconium screens, contains in addition to the characteristic $K\alpha$ radiation a considerable amount of general radiation. Although usually not effective, this general radiation becomes important when the sample being analyzed is composed of crystal grains of certain sizes. The effect under discussion is caused by reflection of this general radiation from the principal atom planes of these crystals. Several experiments, and a geometrical analysis of the positions and orientations of the diffraction effects, confirm this conclusion.

[8] J. O. S. A. and R. S. I. 9, 123-7 (August, 1924).

Contributors to this Issue

Frank Gill, European Chief Engineer of the International Western Electric Company. Mr. Gill has had long experience as a telephone engineer, first, with the United Telephone Company in London, then with the National Telephone Company and later as a consulting engineer. At the outbreak of the war, he was called upon to undertake important work in the Ministry of Munition for which he was later awarded the Order of the British Empire. As European Chief Engineer of the International Western Electric Company, he is taking a leading part in the discussion and study of conditions necessary for the establishment of an adequate long distance telephone service through Europe.

Oliver E. Buckley, B.Sc., Grinnell College, 1909; Ph.D., Cornell University, 1914; Engineering Department, Western Electric Company, 1914–1917; U. S. Army Signal Corps, 1917–1918; Engineering Department, Western Electric Company (Bell Telephone Laboratories), 1918—. During the war Major Buckley had charge of the research section of the Division of Research and Inspection of the Signal Corps, A. E. F. His early work in the Laboratories was concerned principally with the production and measurement of high vacua and with the development of vacuum tubes. More recently he has been connected with the development and applications of magnetic materials and particularly with the development of the permalloy-loaded telegraph cable.

Harvey Fletcher, B.S., Brigham Young, 1907; Ph.D., Chicago, 1911; instructor of physics, Brigham Young, 1907–08; Chicago, 1909–10; Professor, Brigham Young, 1911–16; Engineering Department, Western Electric Company, 1916–24; Bell Telephone Laboratories, Inc., 1925–. During recent years, Dr. Fletcher has conducted extensive investigations in the fields of speech and audition.

Charles W. Carter, Jr., A.B., Harvard, 1920; B.Sc., Oxford, 1923; American Telephone and Telegraph Company, Department of Development and Research, 1923–.

Arthur S. Curtis, Ph.B., 1913; E.E., 1919; Sheffield Scientific School; Instructor in Electrical Engineering, Yale University, 1913–17; Engineering Department, Western Electric Company, 1917–24; Bell Telephone Laboratories, Inc., 1925–. Mr. Curtis' work has been connected with the development of telephone instruments.

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, 1917–24; Bell Telephone Laboratories, Inc., 1925–. Mr. Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

Lloyd Espenschied, Pratt Institute, 1909; United Wireless Telegraph Company as radio operator, summers, 1907–08; Telefunken Wireless Telegraph Company of America, assistant engineer, 1909–10; American Telephone and Telegraph Company, Engineering Department and Department of Development and Research, 1910–. Took part in long distance radio telephone experiments from Washington to Hawaii and Paris, 1915; since then his work has been connected with the development of radio and carrier systems.

Austin Bailey, A.B., University of Kansas, 1915; Ph.D., Cornell University, 1920; assistant and instructor in physics, Cornell, 1915–18; Signal Corps, U. S. A., 1918–19; fellow in physics, Cornell, 1919–20; Corning Glass Works, 1920–21; asst. prof. of physics, University of Kansas, 1921–22; Dept. of Development and Research, 1922–. Dr. Bailey's work while with the American Telephone and Telegraph Company has been largely along the line of methods for making radio transmission measurements.

C. N. Anderson, Ph.B., M.S., University of Wisconsin, 1919; Technical Asst. U. S. Naval forces in France, 1917–19; instructor Engineering Physics, University of Wisconsin, 1919–20; General Electric Co., 1920–21; Fellow to Norway, American-Scandinavian Foundation, 1921–22; American Telephone and Telegraph Co., Dept. of Development and Research, 1922–. Mr. Anderson's work has been chiefly on radio transmission.

readily comprehend the ratios corresponding to any given number of units. Since it is desirable that every unit be based on a decimal system of notation, unless there is some very important reason why it should not, the TU based on the decimal system was chosen. Satisfactory experience during the past year and a half is pointed to as showing the wisdom of having chosen the TU.

A Suspension for Supporting Delicate Instruments.[4] A. L. Johnsrud, Bell Telephone Laboratories, Incorporated, New York. A description, with diagram, is given of a modified Julius suspension designed especially to eliminate disturbances due to vertical vibrations from the building structure. The frame holding the instrument is supported by a system of tape-wound coil springs, which, because of the tightly wound friction tape, damp out mechanical vibrations. The frame with its balancing weights, is heavy (about 120 pounds), and so proportioned in mass that a twisting or tilting impulse, necessary at times in adjusting the instrument, disturbs its moving system only in a secondary degree. This is a second feature of this suspension. Surprisingly effective kinetic insulation is achieved. Quadrant electrometers and a moving magnet galvanometer have remained undisturbed even when heavy trucks were passing on the street seven floors below. This type of suspension, developed some years ago through the efforts of Mr. H. C. Harrison and Mr. J. P. Maxfield, has been adapted for use throughout the Bell Telephone Laboratories in a variety of ways.

Power Amplifiers in Transatlantic Radio Telephony.[5] A. A. Oswald and J. C. Schelleng. The paper describes the development of a 150-kilowatt (output) radio frequency amplifier installation built for transatlantic telephone tests. The characteristics of the single-sideband eliminated-carrier method of transmission are discussed with particular reference to its bearing upon the design of the power apparatus. A classification of amplifiers is proposed in which there are three types distinguished from each other by the particular portion of the tube characteristic used. The water-cooled tubes employed in these tests are briefly described, special consideration being given to their use in a large installation. The system is then shown in outline by means of a block diagram, the elements of which are subsequently discussed in greater detail. The theory, electrical design, and mechanical construction of the last two stages of the amplifier are outlined, including the output and antenna circuits. Means employed to prevent spurious oscillations are described. The method

[4] Journal Opt. Soc. of Am., Vol. X, No. 5, pp. 609–611, May, 1925.

[5] Proc. of I. R. E., Vol. 13, page 313, June, 1925.

used in increasing the transmission band width to a value much greater than that of the antenna is explained. The power requirements of a single sideband installation are outlined and a description of the six-phase rectifier, used as a source of high potential direct current is given, together with a brief theoretical treatment of its operation. Circuit diagrams, photographs, and a number of characteristic curves are discussed.

Production of Single Sideband for Transatlantic Radio Telephony.[6] R. A. HEISING. This paper describes in detail the equipment and circuit used in the production of the single sideband for transatlantic radio telephony in the experiments at Rocky Point. The set consists of two oscillators, two sets of modulators, two filters, and a three-stage amplifier. The oscillators and modulators operate at power levels similar to those in high-frequency communication on land wires. The three-stage amplifier amplifies the sideband produced by these modulators to about a 500-watt level for delivery to the water-cooled tube amplifiers.

The first oscillator operates at about 33,700 cycles. The modulator is balanced to eliminate the carrier; and the first filter selects the lower sideband. In these transatlantic experiments the second oscillator operated at 89,200 cycles, but might operate anywhere between 74,000 and 102,000 cycles. The second modulator, which is also balanced, is supplied with a carrier by the second oscillator and with modulating currents by the first modulator and first filter. The second filter is built to transmit between 41,000 and 71,000 cycles, so that by varying the second oscillator, the resulting sideband, which is the lower sideband produced in the second modulating process, may be placed anywhere between these two figures. Transmission curves for the filters are given as well as some amplitude-frequency performance curves of the set.

A Null-Reading Astatic Magnetometer of Novel Design.[7] RICHARD M. BOZORTH. This instrument is designed for measuring the magnetic properties of very small amounts of material in the form of fine wires, thin tapes, or as thin deposits (electrolytic, evaporated, sputtered) supported on non-magnetic forms. The specimen, 4 cm. long, is mounted parallel to the line joining the two needles, so that its poles produce the maximum torque on the suspended needle system, the position of which is read by mirror and scale. The effect of the magnetizing coil on the needles is annulled once for all by the suitably placing of an auxiliary coil, and the magnetic effect of the

[6] Proc. of I. R. E., Vol. 13, page 291, 1925.

[7] J. O. S. A. and R. S. I. 10, 591-8 (May, 1925).

The Bell System Technical Journal

October, 1925

General Engineering Problems of the Bell System

By H. P. CHARLESWORTH

NOTE: This paper, read before the Bell System Educational Conference, Chicago, June 22–27, 1925, discusses the character and scope of the important problems involved in caring for the growth and operation of the Bell System. The plant extensions necessary to meet service requirements and the necessity of advanced planning are first taken up. The uses of the "Commercial Survey," the "Fundamental Plan" and engineering cost studies are analyzed to illustrate how an engineer attacks the problem of furnishing satisfactory telephone service to the public. A discussion of the New York-Chicago toll cable and the telephone problem in New York City, as illustrative of specific engineering problems, concludes the paper.

THE problem of giving telephone service is quite different from that of most business enterprises. The merchant, for example, may take more business in his store without necessarily always increasing his facilities. The minute we take another subscriber, however, we add to our plant and plant investment. Similarly, in connection with the manufacturing industry, the manufacturer, for instance, is in a position to exercise very direct control over his activities. In the telephone industry, however, our obligation is to take the service as requested and be prepared to deliver it when and as it is required. Furthermore, the activities of the telephone business are of such a nature as to make it essential, regardless of the remoteness of the territory or of the physical and climatic conditions involved, that a way be found, as far as practicable, to construct and maintain the plant and safeguard the service to the public.

To meet these exacting requirements calls for the greatest ingenuity and foresight in the design of the telephone plant and involves careful study of various plans for plant extension and rearrangement with a view to the selection of the most economical and desirable plan. Having determined the fundamentals of design, there must, of course be devised ways and means of safely constructing and efficiently maintaining the plant. Furthermore, as the plant is necessarily scattered over a very large territory and as the different parts must work together satisfactorily and with the most economical results, a high degree of standardization is required, still leaving, however, freedom to adapt the plant to different local conditions. We find evidence on every hand of the value of this standardization, not only during normal conditions, but also during emergencies, when it has been possible to quickly assemble equipment or materials from any part of the system and promptly restore or expand the service as required.

Important engineering problems of great variety, therefore, present themselves on every hand calling for consideration by the engineers in the General Engineering Departments, as well as the Traffic, Plant and Commercial engineers associated with the operating divisions of the companies.

Plant Extensions to Meet Service Requirements

A very large part of the engineering work of the Bell System is concerned with the design of plant extensions to meet expected future service requirements with the maximum economy consistent with maintaining the service standards of the system. I shall not discuss the magnitude of the various activities and requirements of the system, but will recall to your mind a few of the outstanding items to better illustrate the magnitude of this part of the engineering work.

Telephone stations are being connected at the rate of over two and one-quarter million annually.

The resulting net additions or gain in stations per year is approximately 800,000.

To meet this station gain and to replace equipment removed from plant, switchboards are being added at the rate of approximately 1,200,000 station capacity annually.

The Bell System installs in one year approximately 30 billion feet of insulated conductor in lead covered cable ranging in unit sizes from 1 pair to 1,212 pairs. Of this amount, more than 27 billion conductor feet constitute the net annual increase in conductor mileage.

The above plant additions, together with other important items, such as poles, wire, etc., involve a net increase in the telephone plant of nearly three hundred million dollars annually.

It is of interest to note, in this connection, that the annual additions to the telephone plant today are equivalent to the entire plant in service in the Bell System as of about 20 to 25 years ago.

Necessity for Advance Planning

Obviously with a program of this magnitude and of such diversity in the character of its related units, careful advance planning is necessary to insure economical and satisfactory performance.

In the earliest days of the telephone service, the problem of laying out a telephone plant was a simple one. A very small switchboard, simple in character and easily moved, if necessary, was placed in some convenient location, usually in rented quarters, and from that switchboard wires were run one by one as needed, to the premises of those desiring service, either on poles or over house-tops. Under such simple and rudimentary conditions, no serious question of the future needed to be answered. Today, how different is the telephone situation in many large cities, such as Chicago, or throughout the system. Large and specially designed buildings must be constructed for the accommodation of the necessary interconnecting or switching mechanisms; expensive switchboards must be placed in these buildings; conduits must be extended from each of these buildings along appropriate routes to reach the thousands of telephones which receive service from these switchboards; other conduits must be placed between these switchboards and the other buildings and switchboards throughout the city so as to provide the means of intercommunication between the subscribers connected with the switchboards located in different buildings; still other conduits and cables must be placed between these switchboards and the central switchboard or toll board from which radiate cables and conduits and lines extending to the suburban area, to adjacent cities, to all the other principal cities in the United States, and to Canada.

Each of the buildings must be placed in some definite location and it is necessary to plan this well in advance and to direct the growth of the plant toward that location, even though the building may not be built for some years hence. Otherwise, very serious and costly rearrangements of plant would be necessary at the time the office is opened. Furthermore, each building must be planned for some definite ultimate size, although, of course, the whole building need not be built at one time. Ducts cannot be placed under the streets one by one as needed. Public sentiment would not, of course, tolerate the opening of important street routes several times, or even once, each year for the purpose of placing an additional duct. Neither would it be economical, if practicable, to construct conduits in this piecemeal way. The manholes in these conduits must be planned with reference to the number of ducts extending into them, not only the ducts initially placed, but if side runs are to be made from these manholes or if other ducts are to be placed later, this fact must be foreseen and provided for, or extensive and expensive alterations are inevitable at a later date.

I might go on and multiply the conditions which must be met in constructing telephone plant in a country such as ours in which not only the population is growing and moving, but where the demand for telephone service is growing more rapidly than the population. We are in effect planning a growing organism and we must recognize that we are dealing with ultimate tendencies largely beyond control, the effects of which are not capable of exact valuation. However, enough has been said, I believe, to indicate clearly to you that the telephone company on every item of its buildings, conduits and cable construction must constantly answer for itself vital questions as to the future requirements of the system.

This was early recognized, and one of the most important engineering problems of the Bell System has been the formulation of estimates of expected future telephone business both as to quantity and expected location, and the development, from these estimates, of basic plans of procedure, which plans must, of course, be flexible, capable of modification from time to time, and such modifications must be made as changing conditions show them to be advisable.

Our first step in determining the estimated future telephone requirements is to prepare a so-called "Commercial Survey" of the city, covering the requirements fifteen or twenty years ahead. These studies include a critical analysis of the existing market for telephone service, pertinent facts as to the present sale of telephone service, of classes of service and users and forecasts of the market for telephone service at the future date or dates. Consideration is also given to the growth and distribution of population, expected changes in general wage levels, etc., and assumptions of the amount of business that must be sold in each area on the future dates selected under assumed rate conditions.

Having thus determined from the "Commercial Survey" the requirements for telephone service for various parts of the city at the future date assumed, it is next essential to develop a comprehensive plan to serve as a basis for the layout of the plant to meet these requirements. Accordingly, a so-called "Fundamental Plan" is made for the community covering these conditions as estimated fifteen or twenty years hence. The importance of such a plan is obvious, but a brief reference to some of its features will, I believe, be of interest.

In laying out a plan for a city, the engineer might, as an extreme case, center all the subscribers' lines at one building. Obviously, we would have a maximum efficiency in operation in some respects, in that we had grouped all of our switchboards together, but our outside plant costs would be at a maximum and other disadvantages would be experienced. As the other extreme, the engineer might place many small buildings around the city, thus placing the outside plant costs at a minimum, but increasing the difficulty and expense of operating so many centers. Obviously, therefore, there is some arrangement between the two extremes I have cited which would provide the most economical and satisfactory layout of the plant. Several test cases, which in the judgment of the engineer seem promising, are, therefore, studied and the most economical and satisfactory plan determined upon. In completed form, these "Fundamental Plans" furnish us the following essential information upon which to proceed with the more detailed studies covering plant extensions.

a. The number of central office districts which will be required to provide the telephone service most economically and the boundaries of these central office districts.

b. The number of subscribers' lines to be served by each central office district.

c. The proper location for the central office in each district to enable the service to be given most economically with regard to cost of cable plant, land, buildings and other factors.

d. The proper streets and alleys in which to build underground conduit in order to result in a comprehensive, consistent and economical distributing system reaching every city block to be served by underground cable.

e. The most economical number of ducts to provide in each conduit run as it is built.

Our experience has shown that these fundamental plans reduce guesswork to a minimum by utilizing the experience of years in studying questions of telephone growth in order to make careful forecasts on the best possible engineering basis. These fundamental plans, together with related studies, thus provide a general program of plant extension to be followed throughout the period for each of our cities and somewhat similar plans are, of course, undertaken for determining the future requirements of our intercity or toll facilities.

It is evident that both the ultimate arrangement and the program whereby it is to be obtained must have the utmost flexibility in order to meet unforeseen requirements, must work in satisfactorily with the existing plant, which represents an investment of over $2,500,000,000,) must meet immediate service requirements, and also permit full advantage being taken of new developments in the telephone art.

The specific or detailed plan for each project of plant extension, whether within the cities as discussed or between cities in the toll line plant must, of course, be started early enough so that adequate time is allowed for completion of the construction work before the new facilities are required. The complete interval between starting work on such a project and getting it into service can seldom be less than one year and in the case of building and central office equipments must, of course, be longer.

Engineering Cost Studies

Owing to the complexity of the problem of suitable advance planning for the growth in the telephone plant as already discussed, it is evident that in the study of plans for specific projects, selection must generally be made between a choice of arrangements, more than one of which might satisfactorily meet the requirements of the service. It is usually necessary, therefore, that two or more practical plans or programs for construction must be compared so that the most advantageous plan may be selected. An important factor in the selection of all of these cases is a study of the relative economies of the different plans; that is to say, a comparative cost study and as these studies form such an important part of our engineering work, I believe it will be of interest to devote a few moments to a description of the important considerations generally involved.

These engineering cost studies require analysis and consideration of the cost and resulting annual charges for different amounts and types of plant included under each plan. The annual charges comprise items of expense incident to ownership of plant and those that are incurred each year after its installation to keep it in operation and in serviceable condition. As a general thing, in these cost comparisons, another interesting factor is also present; namely, most of the plans which are compared call for expenditures to be made at different periods. For example, one plan might call for erecting a new building at a new location immediately; whereas under the other plan being considered, the necessary additional space required could be secured by adding to an existing building and deferring the complete new project for possibly five or ten years. The relative economy of the plans, therefore, cannot be determined directly by a detailed comparison of the expenditures involved or resulting annual charges, but it is necessary in order to give a fair comparison to express the relative costs of the different plans in terms of present worths, or equivalent annuities which give figures for the total expense in which accurate allowance is made for the variation of expenditures with respect to time.

These engineering cost comparisons may be considered as composed of four parts or operations; namely, the premises or known factors and assumptions; the formulation or set-up of the problem; the solution or mathematical calculations and finally the interpretation of the results. The determination of the premises and formulation of a given problem is, of course, a matter specific to that problem, and here the engineer must exercise sound judgment, for unless the assumptions upon which the work is based are reliable the study itself is of little value. The mathematical calculations are, of course, a definite thing. However, the interpretation of the results must always be a matter of engineering judgment and full weight must be given to those factors which by their nature cannot be evaluated in the cost comparison.

A cost study is a fundamentally important tool in assisting the engineer to reach a decision as to the most desirable plan or program, but as indicated it cannot be used to replace the exercise of judgment on his part. The solution of an engineering problem is, in general, not a matter that can be demonstrated mathematically as can, for example, the proposition, that the square of the hypotenuse of a right triangle is equal to the sum of the square of the two sides. An engineering study rather requires in addition to all of the definite facts that can be brought to bear on the question the exercise of sound judgment on the part of the engineer in weighing the results of the cost study with all related business or other factors bearing on the problem.

Some factors involved in these engineering studies are often of a character which do not permit of expression as a direct charge against a given plan, but must be considered on a broader basis such as the difference in quality or dependability of the service, etc. Also it is important to keep in mind, for example, that, other things being equal, a plan requiring large investments has disadvantages as compared with one requiring a smaller investment so that even though the plan involving a larger investment may prove in from the cost study by a small margin, it may be desirable to adopt the alternative plan so as to avoid tying up considerable amounts of fixed capital. Another question to be kept in mind in interpreting cost studies is whether the more expensive type of plant, usually a higher type of plant, can be adopted satisfactorily at a later date or whether the decision to be made at the present time precludes its adoption later. In the former case it is often wise to go further in deferring fixed capital expenditures than in the latter case. Finally, throughout all of his work the engineer must have foremost in his mind the fact that the telephone system exists for the purpose of furnishing service to the public and the results of his engineering effort should insure a service which is satisfactory from the subscriber's viewpoint.

It is evident from what has been said, I believe, that these engineering cost studies are of great benefit in working out the proper procedure in our engineering work, and I assume they are equally helpful in the engineering of any kind of growing plant. Anything that can reasonably be done, therefore, to give the student an appreciation of the nature, scope, and application of the economic considerations of these engineering problems and to develop his faculties of judgment, imagination, team play, and other related qualities, will doubtless prove of great value to the student in his later engineering work.

Other Phases of Engineering Work

I have thus far described to you some of the very important engineering problems involved in the planning and carrying out of plant extensions to meet expected future service requirements. I would like next to consider with you a few of the engineering problems that present themselves in the actual design or operation of these large extensions to plant as introduced.

The rapid development of the telephone system, including the tremendous growth in the number of telephones in service and the rapid increase in the extent of territory which can be reached from any telephone, has led to a great increase in the importance and difficulty of the technical problems involved in the design and maintenance of the plant.

These technical problems cover a very wide range. The electrical and acoustic problems involved in the transmission of speech have led telephone men to much pioneering work dealing with the flow of sustained and transient alternating currents in electric circuits of all types and in the fundamental nature of speech and hearing itself. Again, the economical design of outside plant with suitable strength and economy involves investigations of characteristics of construction and materials and the preservation of timber, and there are, of course, special mathematical and other problems involved in the design of long cable or wire spans. Buildings and associated central office equipments involve very interesting mechanical and electrical problems in the matter of the layout of the buildings and the arrangement of apparatus to meet exacting requirements. These include many problems in the design of means for automatically supervising the progress of telephone connections and in the design of thousands of types of apparatus to meet specific mechanical and electrical requirements.

What I have already said emphasizes the importance of engineering work involved in the design of new plant. Very interesting engineering studies are, however, also involved in connection with the maintenance of the plant as well. This includes the development of improved maintenance methods and routines and a critical analysis of the results obtained, judged from the points of view of excellency of the service and economy of operation. To use a homely illustration: one might have his automobile completely gone over by a garage every 100 or 200 miles of running with the result that he would probably be reasonably sure of perfect maintenance of the automobile (assuming a perfect garage), but the maintenance costs would be excessively high and out of proportion to the benefit received. On the other hand, however, if no attention is given to the maintenance of the automobile, maintenance costs would be at a minimum but the depreciation would be high, the operation would soon become unsatisfactory and sooner or later the results would be a total interruption to service use. The problem, therefore, evidently is to find the proper balance between overall costs and service results, and this is true, of course, of the various engineering problems to be solved in connection with the maintenance of the telephone plant.

The engineering work of the Bell System also involves, to a large extent, relations with other organizations. These relations are very close with other wire-using companies, including small telephone companies whose lines connect with those of the Bell System. Important relations must be maintained by the engineer with electric power and electric railway companies, as particularly important problems of safety and of service arise due to the proximity between their electric circuits and the telephone circuits. These problems involve provision not only for the protection of the plant and employees against the danger of contact with the wires of other companies but also include coordination of the two systems to prevent excessive inductive effects which often become important where electric power lines or electric railways and telephone lines run parallel to each other. The electric companies and the telephone companies often find it advantageous to enter into arrangements for the joint use of pole lines and this presents many problems requiring consideration by the engineer. It is evident, therefore, that the problems of the telephone engineer cover a very wide and interesting field in mechanical, electrical and other arts, both within the business itself and in relation with other utilities and municipal, state or national bodies or associations.

Specific Projects Illustrating Telephone Engineering Problems

Enough has been said, I believe, in the foregoing to indicate the general nature of the engineering problems handled in the Bell System. It is, of course, impracticable and doubtless would be tiresome in a talk of this character to deal specifically with many detailed engineering problems involved in the work which I have just described in general terms. I believe that you will gather a better appreciation of what some of these problems are from the inspection trips which form an important part of this week's program, than you could by a full discussion of them here. It will probably be of interest, however, before closing to outline briefly one or two typical telephone engineering problems of considerable magnitude.

New York-Chicago Toll Cable

The first large engineering problem I will consider is that relating to the New York-Chicago toll cable as shown in Fig. 1. This cable follows a route from New York through Harrisburg, Pittsburg, Newcastle, Cleveland, and thence to Toledo, and when completed[1] will extend to South Bend and then on to Chicago. For parts of the distance through the congested sections it is underground, and through the open country it is aerial.

Until a comparatively few years ago practically all long toll circuits were in open wire construction; that is, individual wires mounted on separate insulators attached to cross-arms on poles. This was a natural development at first, due to the small number of circuits usually involved, but was also necessary because of the relatively high transmission losses of cable circuits where, as you know, the wires are insulated by wrappings of paper, closely twisted together in pairs and quads, and large numbers of these compressed together within a lead sheath. The rapidly increasing use of toll service, however, pointed to difficulties in providing for future growth with open wire lines. In different parts of the route between Chicago and New York, for example, there were three and four heavily loaded open wire toll lines and the rate of growth was so rapid it was evident that before long difficulty would be experienced in obtaining suitable routes for the additional pole lines required.

Early efforts were accordingly made to devise means which would permit of satisfactory talks through cable and as a result of very intensive research there were developed satisfactory forms of telephone

[1] This cable has recently been completed.

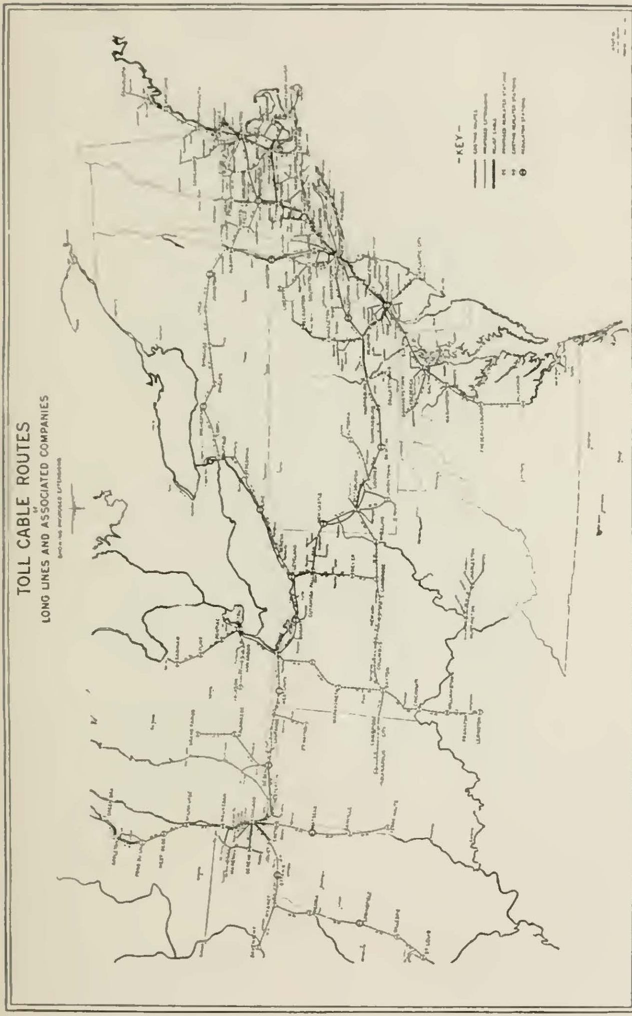


Fig. 1

repeaters; that is, devices for amplifying feeble telephone currents, passing in either direction over a telephone circuit, without appreciable distortion. The most successful repeaters of this type, as you may know, use as the amplifying element the vacuum tube, although the tube itself is but a very small part of the apparatus required for the successful operation of the telephone repeater, and many interesting

Fig. 2—Open wire toll line

engineering problems had to be solved in providing a complete repeater. A full discussion of this very important and interesting development is given in a paper by Mr. Gherardi and Dr. Jewett, published in the Transactions of the A. I. E. E. for 1919.

The toll cable development, based on the use of repeaters as outlined above and many other technical improvements, now makes it possible to give satisfactory service between Chicago and New York and intermediate points over toll cable circuits of such small gauge that close to 300 circuits can be included in a single sheath of 2⅝″ in diameter. The same number of circuits would require four or five very heavily built pole lines of open wire construction such as is shown in Fig. 2.

The construction of the Chicago-New York cable was started in 1918 and will be completed this year. As shown in Fig. 1, the cable is now in service between Chicago and South Bend, Indiana, and between New York and points as far west as Toledo. This cable is one element of a very extensive network of toll cables, particularly in

Fig. 3—Transporting cable reels through Allegheny Mountains

Fig. 4—Toll cable line in Allegheny Mountains

the northeastern part of the country. Important cables in service or being installed out of Chicago, in addition to the New York-Chicago cable, include cables from Chicago to St. Louis, Chicago to Terre Haute, Chicago to Milwaukee, Chicago to Davenport, Iowa. During this year the Bell System is installing over 1,000 miles of toll cable containing more than 2 billion 500 million feet of insulated conductor.

The successful operation of long circuits of this cable network has been brought about only by the solution of very difficult technical problems, some of which have already been mentioned. It may be of interest to state that the long through circuits in this cable will be in the nature of four-wire circuits; in other words, one pair of small gauge wires with repeaters will be used for talking in one direction and

Fig. 5—Typical telephone repeater station

a similar pair so equipped will be used for talking in the other direction. As an illustration of another type of problem involved, it may be of interest to mention that it is necessary to employ automatic regulators which vary with changes in the temperature of the cable conductors, the amplification introduced into the circuit by some of the repeaters. Without regulation, the change in temperature occurring within 24 hours often makes as much as a thousand-fold difference in the amount of electrical energy received over New York-Chicago circuit from the same input, a variation which would, of course, utterly prevent giving service over the circuits.

Aside from the electrical difficulties there were also interesting problems of a mechanical engineering nature to overcome in the desing and placing of the cable, particularly where it passes through the wilderness of the Allegheny Mountains as shown in Figs. 3 and 4.

The cable is for most of its distance strung on pole lines and these lines were designed especially to withstand the stresses caused during sleet storms. The decision as to whether the cable should be underground

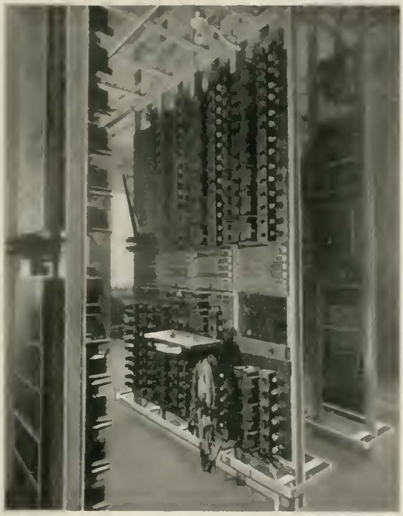


Fig. 6 Bank of 2-wire telephone repeaters

or aerial in the various sections in itself involved many engineering considerations.

In addition to the engineering matters in connection with the cable itself, other interesting problems present themselves, of course, with regard to the design and construction of the telephone repeater stations and their associated equipment, the telephone repeaters being inserted in circuits of this character at intervals of about 50 miles. A typical repeater station is shown in Fig. 5, a bank of two-wire repeaters in Fig. 6, and a bank of four-wire repeaters in Fig. 7.

Fig. 8 shows a view of the completed cable. In this case a loading coil case is also shown, and the picture indicates again the physical problem of erecting a cable through the less accessible sections of the territory. Fig. 9 shows another section of the completed cable through open country, and shows loading coil construction and facilities for

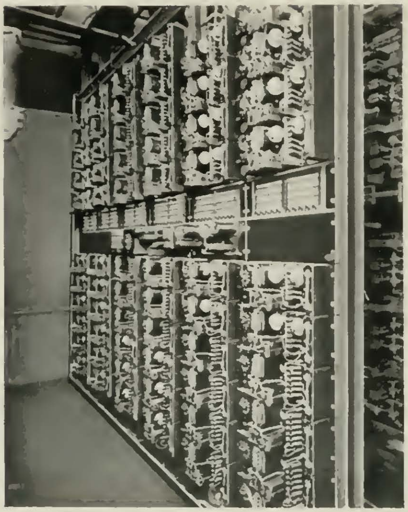


Fig. 7 Bank of 4-wire telephone repeaters

Fig. 8 Toll cable line showing loading coil case

cutting in additional loading coils as required. Fig. 10 gives an interesting view of the cable over the Alleghenies, showing us again the mechanical problems involved in design and construction. In this case the cable follows closely the open wire line, which in time will be dismantled.

It may be of interest in this connection to state that the plans to be compared in the study of toll cable projects generally differ primarily in the dates at which they contemplate supplementing or replacing open wire service by cable. Conditions under which cable becomes economical depends, of course, on many factors. Perhaps the most important single factor is the rate of growth of the circuit requirements. The detailed design of the cable also involves very interesting studies of the economical number of circuits to provide in the cable sheath. Also the economical gauge of each circuit must be considered, comparing in many cases the economies of a larger gauge with those of a smaller gauge provided with a greater number of telephone repeaters.

The design of the toll cable as discussed is but one illustration of the design of the toll plant extension as a whole, a problem which, in general, involves the consideration of the relative desirability of additions to existing open wire toll lines, building new open wire toll lines, applying carrier telephone systems to existing lines or installing toll cable.

Telephone Problem in New York City

As another specific illustration of the telephone engineering problem, I will describe briefly the matter of adequately meeting requirements in a large city, using for purposes of illustration the situation in New York City and the metropolitan area. This particular situation doubtless presents one of the most difficult engineering problems and in some respects is unusual, yet, on the other hand, it fairly represents the kind of engineering problem with which the Bell System engineers must deal at all times.

Fig. 11 indicates clearly the magnitude of the present and future problem in the New York metropolitan area, as viewed from the number of telephones. In 1905 there were 220,000 stations in New York City and 300,000 stations in the metropolitan area. By 1925 the figures had increased to 1,100,000 for New York City and 1,900,000 for the entire area. By 1945 it is estimated there will be over 3,000,000 stations in New York City and over 4,000,000 in the metropolitan area. Part of this growth can be ascribed to the normal increase in the population and part, of course, to the tendency to make more use of

Fig. 9—Toll cable line through open country

Fig. 10—Cable and open wire toll line in Allegheny Mountains

the telephone. In addition, part of the growth is due to the conditions following the World War and the general economic trend.

Comparing 1924 with 1914, wholesale commodity prices, as you know, have risen over 50 per cent; the cost of living over 60 per cent; wages in manufacturing industries over 100 per cent, while in the same period telephone rates generally have increased less than 30 per cent.

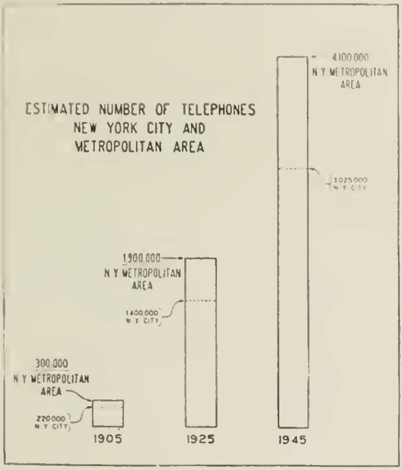


Fig. 11

and even less than this in some of the larger cities. Telephone service, therefore, represents a large value for its price and in a situation like Greater New York City, where there are between seven and eight million people, it is but natural that the new situation in the economic balance of things, together with the low price of service shown, would make for a very substantial increase in the demand for telephone service. This has, of course, also been true elsewhere.

As I have shown there are at present a total of over one million telephone stations within New York City proper served from about 130 central offices, 26 offices having been added last year. The predictions are that within the next twenty years the stations and central offices will have more than doubled. Each subscriber in this great network must be able to reach promptly every other subscriber.

Due to the large area involved, a great number of calls within the city necessitates extra charges, which means that they must be specially supervised and recorded. There are many different classes of service furnished the public, such as measured rate, flat rate, coinbox, etc., and, of course, such other special services as Information service. Not only individual lines but party lines and private exchanges must be cared for. Furthermore, the demands for service to the extensive area surrounding this great city, as well as the large number of cities, towns and rural communities throughout the entire country, require that provision be made for thousands of toll messages daily. The problem of giving satisfactory service under these conditions and under the complications that come with the tremendous growth referred to is a very important one and requires careful and constant study.

In order to properly care for this complex problem of furnishing telephone service in large cities, telephone engineers in line with the efforts which have been made from the time of the early switchboards have endeavored to perform the various operations automatically so far as consistent with service requirements. While the switchboards which you saw yesterday are called "manual" switchboards, you doubtless noted from the demonstration and your visit through the central office that many of the operating features are automatic in character. The latest step in this general trend of development has been to develop a switchboard which would provide for completing many classes of calls entirely without the aid of an operator, and these new machine switching equipments which you will see today are gradually being introduced into New York, Chicago, and other large cities. This is a large problem in itself and involves not only the completion of calls from machine switching subscribers to other machine switching subscribers, but the completion of calls incoming to machine switching offices from manual offices and outgoing to manual offices. This must be done without reaction on the service or inconvenience to the subscribers and so that the machine equipment and the manually operated switchboards will work together as a coordinated whole.

I do not know of any mechanical device that reminds one so much of the functioning of the human brain as does this mechanism for completing calls following the dialing operation. The completion of a simple call, while quite involved in itself, is by no means the complete problem. There must be a great many other features provided, such, for one example, as where a register is provided on the subscriber's line to register the number of calls under measured rate service. In these cases it is necessary to insure that there shall be proper registration by the machine and the mechanism is so arranged, therefore, that on the completion of the call it will test the line to make sure that everything was normal before registration is actually performed. Similarly, all the way through the completion of the regular and special classes of calls it is necessary for the mechanism to perform just such intricate functions as that described.

The engineering of the interoffice trunk layout in a city like New York is also an important and interesting problem, not only because of its magnitude but because of the almost unlimited variations which

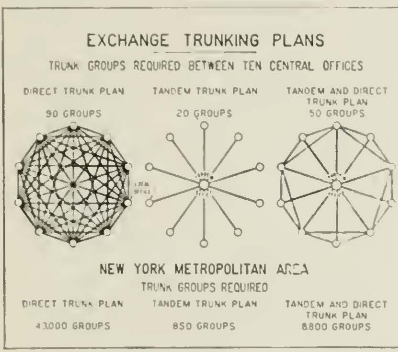


Fig. 12

might be employed, a large number of which must be carefully considered in connection with additions to the plant. In opening new central offices, trunk circuits must be provided between each new office and the existing offices and also between the new offices themselves.

Fig. 12 illustrates the range of trunking layouts which might be used. With the 10 offices assumed and direct trunks between each office and every other office, 90 groups of trunks would be required. With the so-called full tandem operation; that is, under an arrangement whereby each office reaches every other office through a central point, 20 groups of trunks would be required. Between these two extremes with some offices reaching certain other offices through the tandem center and certain others by direct trunks, a great many combinations would be possible. In the case assumed 50 groups appeared to be the best combination. The data given at the bottom of Fig. 12 are of particular interest in this connection. As will be noted, if only direct trunks were employed in the metropolitan area, some 43,000 groups would be required. On the other hand, if we followed only the strictly tandem plan, 850 groups would be required but as previously indicated, unwarranted switching costs would be

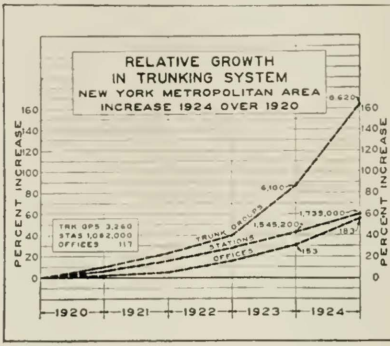


Fig. 13

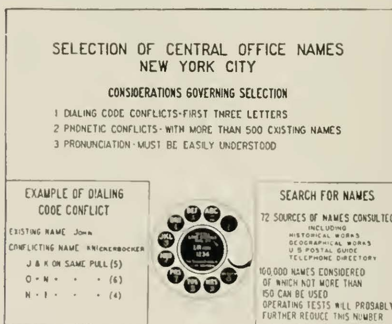


Fig. 14

involved. By establishing a plan, however, involving both tandem and direct trunks, the most economical plan can be determined upon and in this case about 9,000 groups of trunks are required. Fig. 13 shows how rapidly the trunk groups increase with the addition of stations and central offices. You can well imagine the engineering problem involved in working out the most efficient trunking plan for a city such as New York or Chicago.

Aside from the layout of the trunk plant itself, the engineering work involves the design and construction of the underground subway system and the design of the physical cable plant. In one year in

Fig. 15 Bowling Green telephone building, New York City

New York City alone, enough cable has been installed and placed in service to make a cable containing 1,200 wires reaching from New York to Chicago.

The expansion of the metropolitan plant to care for the increase in the number of subscribers also involves, of course, opening many new offices and the provision of new switchboards and additions to the existing switchboards. The matter of selecting the name for a new central office would at first appear to be a simple one, but as indicated by Fig. 14 it is a very involved problem in itself. As will be noted, there are many questions to be considered. One feature relates to the matter of dialing. It is interesting to note from Fig. 14, however, that while the name "John" does not seem in any way to conflict with the name "Knickerbocker," yet these two names could not be used together in the same city because of conflict in the dialing process. Phonetic conflicts are also exceedingly important in telephone operation. In fact, they form one of the most important factors that must be considered in the selection of an office name. Pronunciation of the name must also be easily understood. Thus we find that in the case of the metropolitan area something like 72 sources of names were consulted; for instance, historical works, geographical works, postal

Fig. 16—West 36th Street building, New York City

guides, telephone directories, and other sources, and out of 100,000 names considered not more than 150 could be used and possibly some of these on further study will have to be eliminated. I have mentioned this detail of operation simply to illustrate the variety of the problems for the telephone engineer and the extent to which he must consider them in order to insure the grade of service we are all striving for.

The erection of new buildings and additions to existing buildings is also a large problem, there being 12 new buildings and 21 additions erected in New York during 1923 and 1924. It might be interesting to note that for these buildings and equipments it is necessary to consider not only the proper association of the various elements of the central office unit from the viewpoint of securing satisfactory operation and maintenance conditions, but also to provide for an orderly growth of the different parts of equipment and building. Further, the central office layout must be considered from the point of view of costs which

Fig. 17—Long distance telephone building, New York City

may vary over a wide range under the different arrangements which might be used. This you will better appreciate from your visits through the offices.

I will next show you a few cases which will illustrate some of the problems in the way of providing building space to house switchboard equipments in these large metropolitan areas.

Fig. 15 is a photograph of the Bowling Green building, located in the extreme lower end of Manhattan Island and which will provide space for switchboard requirements for that part of New York City.

Fig. 16 gives a rather interesting example of another of the large New York telephone buildings, this case being the one located in West 36th Street in the neighborhood of the Pennsylvania Station. This building and equipment involve an expenditure of $15,000,000 and is equipped to serve over 100,000 stations. In other words, we find in this one building and the associated switchboards on subscriber's premises, provision for handling more stations, for example, than are in service in a city the size of Baltimore, with a population of nearly

Fig. 18—Barclay-Vesey telephone building under construction, New York City

800,000, giving you a further idea of the problem of providing service in these large metropolitan centers.

Fig. 17 illustrates the building in New York devoted to the centering of all long distance lines. Facilities are also provided for connecting together the various offices of the city for switching to suburban points through one of those tandem boards of which I spoke, as well as for switching to the great network of toll lines running out to all important points throughout the country. While there are some local switchboard facilities in this building, practically all the space is devoted to handling toll traffic.

Fig. 18 illustrates the new building being built for the New York Telephone Company on West Street in the lower part of Manhattan. This building is designed to house a large number of units of machine switching equipment, and the upper part will be utilized for the administrative offices of the Company. This further illustrates the type of building required in these large centers, and the many engineering problems involved.

I might go on at length, giving one problem after another, by way of illustration, but I think enough has been said to give you a general idea of the nature and great variety of the telephone engineering problem involving, as it does, almost every phase of the mechanical, electrical, and other arts. It is obviously necessary for the engineer not only to consider the technical problems involved in each of these matters, but to a greater extent it seems to me than almost any other situation I have encountered, it is necessary for him to take into account all of the related broad operating and business factors which are naturally to be found in an industry of the magnitude of the Bell System.

Engineering Planning for Manufacture[1]

By G. A. PENNOCK

Synopsis: This article discusses the complete analysis, from a manufacturing point of view, to which every item of telephone apparatus is submitted at the Hawthorne Plant of the Western Electric Company. These works employing, at present, about 25,000, produce over 110,000 different kinds of parts which enter into some 13,000 separate forms of apparatus. The advantages of careful engineering analysis of each new job coming to the factory, as well as those which have been in production, are brought out. The various steps which are worked out in connection with each analysis are as follows: manufacturing drawings; the proper manufacturing operations and their sequence; the machines best adapted to carrying out these operations; determination of the kind of tools, gauges, weighing and other equipment; the determination of the probable hourly output for each operation; the grade and rate of pay for the operators; the kind and amount of raw material required; manufacturing layouts which tell the entire shop organization; each step in the production of the parts, and finally the best rate to be paid for each operation. In conclusion, the author discusses the personnel of the Planning Organization.

Introduction

THE essence of the successful operation of any industrial establishment is contained in the maxim "Plan your work—then work your plan." The first part of this maxim is by far the most important since the ability to work any plan depends fundamentally upon the excellence of the plan itself.

Farsighted planning, as applied to elementary factory operations, is a relatively simple problem. For example, the problem involved in planning the work of a foundry is to a great extent merely the the duplication of plans already standardized, but in a plant manufacturing widely diversified products, such as we have at Hawthorne, planning becomes at once more difficult and essential.

The General Manufacturing Department of the Western Electric Company provides the Bell System with telephone equipment which involves the production of over 13,000 separate and distinct forms of apparatus, in the construction of which there are used over 110,000 different kinds of parts made from 18,000 different kinds, sizes, and shapes of raw material. A number of these parts are produced in very small quantities.

The production of the varied product mentioned above involves not only all the usual wood and metal working operations, but also such lines of manufacture as:- glass making, textile dyeing, manufacture of porcelain, electrolytic iron, vulcanized and phenolized fibre,

[1] Paper read before the Bell System Educational Conference, Chicago, June 22-27, 1925.

soft and hard rubber in the form of sheet, rod, tube, and molded shapes, the insulation of wire with textiles, enamels, and paper, and the conversion of copper billets into wire.

These materials are used for making parts which, generally speaking, are quite small in size when compared with parts used in steam locomotives, gas engines, dynamos, and other kindred equipment common to the electrical and mechanical fields.

The fact that the parts are small in dimension, however, does not mean that the manufacturing difficulties are in proportion. On the contrary the problems involved in their manufacture are often times in an inverse ratio to the size of the part.

Fig. 1 shows a crank shaft about three feet long and the shaft used in the calling dial for machine switching about an inch and a half long. The layout of the operations required for machining the crank shaft is shown in the upper left hand corner. There are a total of eight.

Below at the left is shown the layout of operations for making the shaft for the dial. There are a total of eighteen.

As you will note from the data at the right, the number of machines involved is, roughly, the same in each case. These data illustrate the fact, however, that the small part may be more complicated and involve more engineering problems that the larger part.

Planning for the Future

As the manufacturing unit of the Bell System, the Western Electric Company in planning its production has had to bear in mind, first, that the facilities shall be adequate to turn out the tremendous volume of apparatus and equipment required from year to year; second, that the System's supply of equipment must be planned to eliminate, so far as is humanly possible, any interruptions; and third, that the System must get its equipment at the lowest possible cost.

Briefly, our program for providing buildings and equipment for the future is based on a five-year forecast of business made by each Associate Company and summarized by the American Telephone and Telegraph Company.

It takes approximately two years to erect and equip new buildings; consequently, capacity studies on floor space are made two years or more in advance and tool and machine equipment studies are made one year or more in advance, as this equipment can usually be provided in one year.

NOTE COMPARISON OF SIZE OF SHAFTS

SHAFT FOR No. 2 TYPE CALLING DIAL

Oper's. Req'd. (18)

Rough form thread portion, and O. D. counterbore, finish turn, thread and cut off
  Limits ±.0015" for diam.
  ±.002" for l'gth.

Rough and fin. form 2 diams, shear 1 diam., thd. and polish.
  Limits +.003",-.005" l'gth.

Shear S. C. face to l'ght. burr and polish long end. +.000"
  Limits −.001" for diam.

Straddle mill flats—mill four (4) slots.
(2) oper's. +.000"
  Limits −.002".

Machine and Tools

Davenport Auto Screw Machine, 5 chucks, 5 plain and form tools, 1 thread die and three gauges.

No. 1 B. & S. H. S. M.

1 chuck, 2 form tools, thread die, emery stick and 4 gauges

No. 1 B. & S. H. S. M.

1 chuck, 2 form tools, emery stick and 5 gauges

Hand Mill

Milling fixture, vise and jaws and spec. cutters

CRANKSHAFT FOR No. 21 BLISS PUNCH PRESS

Oper's. Req'd. (8)

Rough and finish, center, turn face complete and polish limits diam. +.008",-.000" length ±.008". Mill concave keyways, 3/4" slot and 4 flats—3 oper's. Limits ±.005". Drill 3/4" x 3/8" hole.

Machines and Tools

Hendy 12" x 5' lathe, center drill, turning and polishing tools. No. 3 B. & S. mill machine milling fixtures, arbors and cutters. Cincinnati 1 sp. D.P. drill jig and drill

Fig. 1—Dial Shaft vs. Crank Shaft

The Advantages of Planning

In order to meet the requirements of the telephone business, the Engineering Departments of the System are constantly developing new designs and changing present designs with the object of improving the quality of, or reducing the cost of telephone service. This means that the products that we are manufacturing are constantly undergoing development, with the result that we are continually confronted with changing manufacturing problems.

The decisions reached by the various organizations of the Bell System to proceed with the introduction of the new and changed designs just referred to are based entirely on improved service, lower costs, or both; consequently, before any work on new developments can be done the Manufacturing Department must furnish firm estimates of the cost of one or any number of pieces of apparatus that may be required.

This is made possible by our ability to plan a job in detail on paper and to make an accurate appraisal of the manufacturing costs before production is started. The cost established, selling prices can be determined, and a final decision made by the System as to the merits of any new devlopment.

Furthermore, by scrutinizing the design and concentrating on the various manufacturing operations to be used before the tools are built, numerous changes can be introduced to facilitate manufacture and in this way avoid getting into the factory what have been termed "hospital jobs" which result in retarded production and inflated costs.

The two designs illustrated in Fig. 2 bring out what is possible in a manufacturing analysis of an engineering design. The part shown is the mounting plate used in the calling dial. The design originally showed ears, which were blanked out and turned over toward the inside of the blank and perforated, as shown in the upper view.

The lower view shows the design as it was developed due to the Manufacturing Department's suggestions to blank the ears from the inside of the blank and turn them outward, thus locating the mounting holes in exactly the same position as the engineering design, but saving material. It also simplified the bending of the ears. Instead of a double bend, there is an S bend. The holes were made larger also to permit perforating instead of drilling. The lugs and holes were also unevenly spaced so as to make it impossible to perforate or assemble the part in the wrong position. In shop language, the part was made "fool proof" in this respect, whereas the model was not.

It was formerly common practice among many manufacturers to leave the actual planning of the job to the shop foreman and to some extent this practice still exists. Obviously, under this plan, only the more commonly known methods will be employed as the shop man is not in a position to avail himself of the mass of engineering knowledge that has accumulated in connection with such work. We are convinced that the returns from engineering the actual manufacturing operations are as great as those realized from engineering the design of the product.

CALLING DIAL NUMBER
PLATE SUPPORT

ORIGINAL PROPOSED DESIGN
Objectionable features—lugs formed inward, requires large blank and a cam action tool or two operations in forming. Small holes do not permit perforating

DESIGN FINALLY ADOPTED
Results of comments—lugs formed outward decreases the size of blank permits combined embossing and forming in simple tool. Holes increased in size

Fig. 2—A Modified Part

Factory Arrangement

Before describing our planning work more in detail, a few words should be said about our arrangement of machine equipment. Our metal working machine departments are laid out in such manner that the manufacturing operations are grouped into departments by class of work or operation and not by class of product. Each department performs some definite kind of operation, and each handles all the parts that require that particular operation. Thus we have punch press departments, screw machine departments, a milling department, a drilling department, etc.

The parts produced in these specialized departments pass in proper sequence through all the departments that have work to do on them and finally reach the assembly departments, where they are made up into finished units of apparatus.

The advantages of this method of dividing manufacturing work are that it minimizes investment by avoiding duplication, increases machine activity, provides greater flexibility of equipment, and permits the training of unskilled labor to the point of full productivity in the shortest time.

The conclusion may have been reached that departmental groupings by classes of machines such as have been described is all right for a business of little variety, but that in such a large endeavor handling so diversified a product, it would seem nearly impossible to maintain a proper balance of equipment in all the departments.

As a matter of fact, adjustments are frequently made due to increased or decreased demands, and we frequently have to step up or down both our rate of production and our capacity for certain lines of products or certain definite articles.

To meet this situation, we have capacity data giving the number of hours required by machine operations, assembly operations, etc., for one thousand pieces of each kind of apparatus. With this information, we can readily compute the increase or decrease in shop equipment due to changes in schedules.

There are, of course, some departures from this general practice of functionalizing our machine departments in the case of certain products that require a large amount of special machinery. In these cases, the few "general use" type machines required are grouped with the special machinery into a department for the complete manufacture of the article.

This special practice is also carried out in connection with the manufacture of certain piece parts. These cases are confined to a few parts manufactured in large quantities where it is found expedient to group a variety of machines in order to reduce the amount of handling to a minimum. An example of this is the manufacture of the top part of the desk stand which supports the transmitter, which we know as the "lug holder." This part is made from brass tubing. The operations involved in making the part are, cut to length, burr, several swaging operations, and a number of punch press operations, such as perforating, embossing, and trimming. We have in this case grouped together in the proper sequence the required number and sizes of milling, burring, swaging and hammering machines and punch presses.

Jobbing Shop

We also have a group of departments known collectively as the "Jobbing Shop" which is equipped to perform all the usual machining operations. These departments handle the manufacture of special apparatus, which is made in such small quantities that it does not pay to make the elaborate manufacturing preparations which are justifiable in the case of heavy running apparatus for which there is an established demand.

To give you some picture of just what we set out to do when we plan a job, the following different steps or problems which must be worked out are enumerated briefly:

1st. Manufacturing Drawings.
These drawings tell the shop in detail what is to be made and what the requirements are.

2nd. Manufacturing Operations.
The actual operations required to produce the parts and their proper sequence are decided upon.

3rd. The machines on which the operations are to be performed are determined.

4th. The kind of tools, fixtures, gauges, conveying, and other equipment to be used is determined.

5th. An expected hourly output for each operation is set up.

6th. The grade and rate of pay of the operators to be employed are determined.

7th. The kind and amount of raw material required per thousand parts and the form in which it shall be purchased are determined.

8th. Manufacturing Layouts.
These layouts tell the entire shop organization each step in making the parts shown on the manufacturing drawings.

9th. The piece rate to be paid for each operation is determined after actual manufacture is started.

Manufacturing Drawings

The manufacturing drawings prepared for any piece of Western Electric apparatus comprise complete detail drawings for each part, an assembly drawing showing how the various parts are associated, a stock list of the parts required and the quantities of each, and a test sheet which shows the mechanical and electrical requirements which the apparatus must meet in order to insure satisfactory performance in the System.

In the preparation of these drawings, standards are followed which insure that the designs as far as possible will permit of rugged tool construction which will insure long tool life; that the holes are of such dimensions as will permit them to be perforated wherever possible; that thread sizes for the tapped holes selected are such as to insure minimum tap breakage; and other similar details.

Manufacturing Operations

Before deciding upon the manufacturing operations for any part, a careful detailed analysis is made by the Planning Engineers to determine just what operations are required and how the operations shall be performed in order to obtain a satisfactory production in the most economical way.

In the case of simple parts, it is not a difficult task to determine the manufacturing operations required and their proper sequence. A large proportion of the parts, however, is in the fairly difficult class, and the ingenuity of the Planning Engineer is called upon, together with the advice and guidance of his superiors, in determining the manufacturing operations to be used in these cases.

A fair proportion of our product makes up what might be called the "difficult class" of parts to manufacture, and in setting up the proper procedure in these cases, we frequently hold conferences where the best talent along the particular lines under consideration is called into consultation in determining the best procedure. In many of these cases actual experimentation is carried on before the final tool line-up is decided upon.

Machine Equipment

The machine equipment on which the operations are to be performed is the next thing given consideration, and the most important features are:

1st. To select a machine that is capable of producing the parts to the desired accuracy.

2nd. To select a machine that will result in the maximum production, keeping in mind, of course, the accuracy required.

3rd. To give proper consideration to the investment, maintenance and overhead charges incurred by the machine selected so that these charges do not offset production economies expected.

4th. To insure that the machine selected is up to date with regard to the latest machine practice developments worked out by Hawthorne and by commercial machine manufacturers.

There are, of course, many other features which must be taken into account in selecting the machines for the manufacture of various kinds of parts.

In the case of blanking operation on a punch press, the object is to secure the smallest and therefore the fastest press which has sufficient tonnage capacity to perform the operation required.

Where the part is to be manufactured on an automatic screw machine, the problem is to select the fastest machine that will produce the work to the accuracy required, and at the same time select a machine that has a sufficient number of spindles and tool positions to permit all the operations required being performed before the parts are finally cut from the rod.

A part having a large number of holes to be drilled will necessitate the selection of a multiple spindle machine that can be set up to produce the maximum number of holes in one or several parts at each operation of the drill press.

We have worked out numerous improvements in commercial machinery that have now been incorporated in the product of many machine tool manufacturers. Some of the most important of these are motor driven punch presses, screw machines, milling machines, lathes, etc.

Fig. 3 shows the old belt-driven milling machine. It does not give you a true picture of the whole job, since the overhead drive which is the most objectionable feature does not show in the picture.

Fig. 3 Belt Driven Milling Machine

Fig. 4 shows the modern motor-driven milling machine with the motor mounted in the base and a chain drive enclosed in the housing at the back driving the spindle.

At our suggestion, several of the largest manufacturers of screw machines have incorporated screw slotting devices as standard equipment for multiple spindle machines.

We have just recently worked out a design whereby a high speed screw machine, formerly adapted to brass parts only, can now have its spindle speed reduced through change gears so as to make it adaptable for iron and nickel silver parts, thus providing greater flexibility.

Punch presses were formerly liable to serious damage if two blanks were accidentally placed in a forming die. We have worked out a design of ram which contains a "shear ring." This consists of a soft metal ring so incorporated in the connecting rod of the press as to

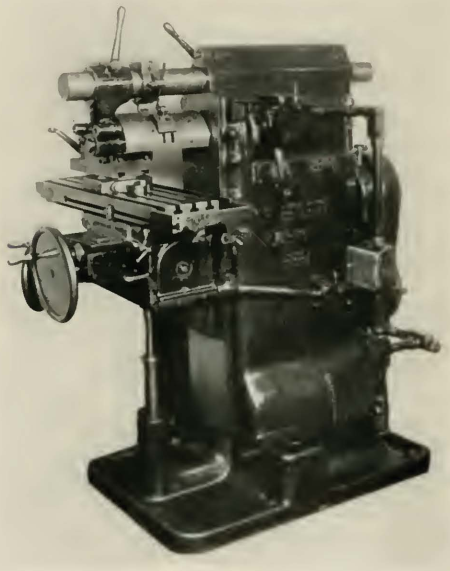


Fig. 4—Motor Driven Milling Machine

shear at any predetermined pressure, thus allowing the connecting rod to telescope instead of breaking the die or frame of the press. This improvement permits operating punch presses safely at greater speeds than are usual on this type of equipment.

Numerous other similar improvements have been worked out, many of which have been patented.

Tools

The annual demand for the product is the most important factor in determining the kind of tool, fixture, and gauge equipment to be provided.

Our most intricate engineering problems arise in connection with punch press tools as there is almost no limit to the variety of operations that can be performed on this type of machine.

If the demand for a part made on a punch press is small, it is often found more economical to build simple tools which will blank out, perforate and form in separate operations, rather than to build more elaborate tools at a higher cost which will combine two or more operations into one.

The effect of quantity on the design of tools may best be shown by a concrete case.

THE EFFECT OF ANNUAL DEMAND ON CHOICE OF MANUFACTURING METHOD

Yearly Req.	Material	Type of Machine	Type of Tool	Tool Cost	Cost per M	Tool Cost per M Parts	Saving per Year
5,000	5/8" Brass Rod	Hand Screw Machine	General use Tools	. . .	$10.00		
30,000	1/16" Sheet	Punch Press	1 at a time Tandem Perforating and Blanking	$150.00	2.30	$5.00	$231.00
500,000	" "	" "	3 at a time Tandem Perforating Perforating and Blanking	400.00	1.72	.80	290.00
3,000,000	" "	" "	7 at a time Tandem Perforating and Blanking	600.00	1.52	.20	600.00

Fig. 5

Take, for illustration, the case of a simple brass washer 5/8" in diameter, 1/16" thick and having a 1/4" hole. As shown in Fig. 5, with a requirement of 5,000 a year, the washer would be made from rod stock in a hand screw machine using general use tools at a cost of $10.00 a thousand; for 30,000 a year it would be made from sheet stock in a punch press using a one-at-a-time tool, at a cost of $2.30 a thousand; for 500,000 a year a three-at-a-time tool would be used at a cost of $1.72 a thousand; for 3,000,000 a year a seven-at-a-time tool would be used at a cost of $1.52 a thousand.

In each one of these steps, as shown in the columns at the right, the additional tool investment, necessitated by the more advanced method, would be liquidated in one year by the decreased manufacturing cost.

Where a high degree of accuracy is required on a piece of apparatus, the overall effect on the tool equipment is to require a greater number of individual tools, as well as to require tools of a higher grade of workmanship. For instance, it may be necessary in the case of a punch press part to hold certain dimensions of the blank to extremely close limits, and this quite often requires an additional operation of shaving the blank to size. This adds an additional tool to the equipment, as well as requiring a tool of greater accuracy.

You will appreciate that the matter of interchangeability is one of great importance—first, because the parts must go together in the assembly departments without any further fitting—and second, the parts and pieces of apparatus shipped over the entire country for repairs and maintenance must be exact duplicates of the old.

It costs more to make interchangeable parts than to make inaccurate ones that are not always interchangeable, and the Planning Engineer can control the tool and manufacturing costs very largely by his judgment in the selection of limits.

Hourly Output

The Planning Engineer, in analyzing the work on a given part for the operations, machines and tools to be provided, from his experience and training in the particular kind of work he is handling, is able to establish an expected hourly production for each operation he handles. He is, of course, guided in this by his experience on similar parts and by the speed of the machines selected for the operation.

The setting up of the expected output for assembling operations is more difficult, but here also the special training and experience of the engineer along that line of assembly work enable him to set up an expected output which is approximately accurate. In some cases, we go so far as to tear down and reassemble models of the apparatus in order to obtain the necessary data.

The output per hour on each operation enables the engineer to compute the number of each kind of tool, including spares which must be built, to produce the required quantity of each part. The number of tools required is obviously dependent on the speed of the operation, and here again you see the effect on tool costs if the engineer fails to select the fastest machine suitable. When it is considered that we have nearly $3,000,000 invested in tools for the manufacture of panel machine switching apparatus alone, it can be appreciated what planning means to us.

Labor Grade

The Planning Engineer, in addition to establishing the values already mentioned, has also the responsibility of selecting the grade of labor which is to be used in performing the various operations. Different grades have been established for men and women, and each grade covers a sufficiently broad range of rates to enable us to hire the employees at the starting rate of the grade and to advance them in the grade as they become more proficient and experienced.

Raw Material

The Planning Engineer specifies the kind and amount of raw material required for each part including the scrap allowance. He also specifies the form in which it shall be purchased—that is to say, whether in rod, tubing, and in the case of sheet stock whether in the shape of sheets, strips, or rolls.

Manufacturing Layouts

The next step in preparing a piece of apparatus for manufacture is the working up of detailed manufacturing layouts. These layouts constitute the "sailing orders" for the shop, covering each operation to be performed, how the work is to be done, the sequence of the operations, the tools and machinery to be used, raw material and quantity required, and the stock room to which the parts shall be delivered upon completion.

These layouts are got out in the form of duplicated sheets and a complete layout for each part is sent to every department having work to perform.

Piece Rates

When all the preparation steps have been completed and after the various operations have been tried out and are running in the operating departments on a satisfactory commercial basis, the Planning Organization proceeds to establish piece rates on each operation.

The piece rates are established by the same organization of engineers who plan the work, and the responsibility of seeing that the estimated outputs are realized devolves upon this organization. Before proceeding with the studies involved in establishing the piece rate, the Planning Engineer checks back against the original planning data and the manufacturing layout, and, in this way, ascertains the method as originally laid out, together with the expected outputs. His task then becomes one of seeing that the expected output or better is attained.

This, in many cases, involves a very detailed time and motion study of the elementary operations necessary to complete the job in order that it be brought to a high state of efficiency. In cases where the expected output cannot be realized by the original method, other methods are worked out wherever possible to bring about the desired result.

Just a word right here on our piece rate policy: when piece work was introduced many years ago, the policy was established that after a rate had been once issued it should not be cut unless a change had been made in the method of manufacture. In other words, we take the stand that an issued rate is a contract which cannot be revoked so long as the operation is done in the same manner as covered by the piece rate card.

To satisfactorily carry out a policy of this kind, it is obvious that our piece rate setting must be something more than mere stop watch observation. In order that piece rates are established which are accurate and fair to both the employee and the Company, it is necessary that the engineers setting the rates be well versed in the class of work being rated, and have a thorough knowledge of the amount of work which can be consistently produced by the operators.

Our experience with the straight piece work form of incentive has been very gratifying, and in our opinion this is very largely due to the following three reasons:

1st. Our policy of not cutting rates.

2nd. Our practice of making careful time studies in setting our our rates.

3rd. Our guaranteeing the employee's day rate regardless of his earnings on the piece rate.

The work of the Planning Engineer is not completed, however, upon the establishment of the piece rate, since it still rests with him to clear any difficulties the shop may experience due to any shortcomings of any of the planning work.

If the raw material provided will not satisfactorily produce the parts, he is called upon either to add operations or to specify other material; if the tools will not produce the parts to the required accuracy, or at the required rate, he is called upon to have satisfactory changes made to the tools or to provide new equipment

In case the operators are unable to produce sufficient parts to make satisfactory piece work earnings after a reasonable trial, the Planning Engineer is called upon to either demonstrate that satisfactory earnings can be made, or to increase the rate.

The Planning Engineer is also called upon to assist in overcoming manufacturing difficulties for which he is not directly responsible, and a special unit has been set up to assist the shop in cases of this kind when difficulties are encountered.

From this, it can be seen that the Planning Engineer has not only the responsibility of planning the work, but he is also charged with seeing to it that the plan works out.

Cost Reduction Work

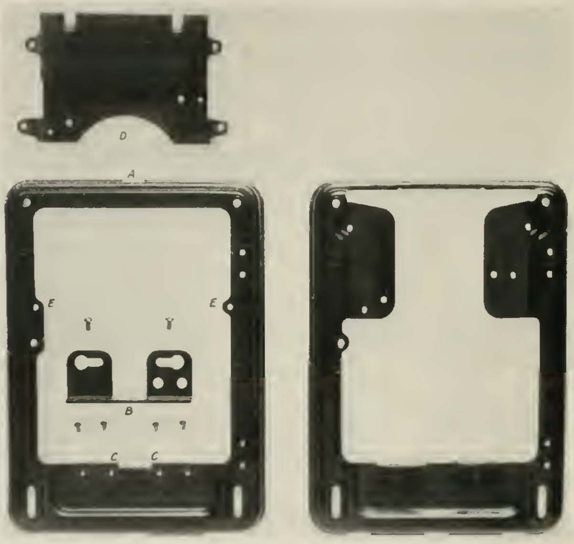
There is still one more highly important function performed by our Planning Organization, viz., Cost Reduction Work.

It might appear that after the careful thought already given to the methods to be employed in producing a piece of apparatus, the necessity for further study has been eliminated. This, however, is not the case, since in the original planning we must adhere closely to methods and processes that have been proved in, in order that the products may be produced on a specified date and at a predetermined cost.

In other words, we cannot take any short cuts at this stage of the work that we are not sure will work out successfully. However, after the piece of apparatus is in production, we are in a position to review the case and try out new ideas, improved methods, tools and machinery, without jeopardizing production. Naturally, any improvements worked out successfully by the Cost Reduction Engineers are later used by the regular Planning Organization when applicable on future work.

This cost reduction work is handled on a strictly business basis, i.e., the cost of the case is charged up against the savings effected and our records show that the returns on this work are very high.

There is a typical illustration of a cost reduction case shown in Fig. 6. This is the base for the sub set housing on which the apparatus is mounted. The old design is shown at the left. There were three separate pieces which had to be assembled together. Part B was riveted to the base A at *c*, *c* to form the ears which stand at right angles to the base. Part D was assembled to the base *A* with two machine screws, *E*, *E*. The design was changed at the suggestion of the manufacturing department to make the part in one piece. It had previously been thought too complicated to combine all these operations in one part, but the tools were successfully developed, and the saving on this particular job amounted to something like one hundred thousand dollars a year, or about ten cents a piece.



ORIGINAL DESIGN
Consists of three individual parts, requires assembly with rivets and screws

ADOPTED DESIGN
One piece construction. Same No. of operations required to make as body of old design. No brackets required.

Fig. 6—Sub Set Base

THE PERSONNEL

So far the job we have to do and how we do it has only been dealt with, and the qualifications and training of the personnel required have not been mentioned.

Our Planning Organization is laid out in a manner similar to our shop departments; that is, the planning of the various manufacturing operations is divided into class of work or operations and not by class of product, each class being handled by a group of planning engineers in charge of an expert thoroughly familiar with the line of manufacture he handles. In this manner each group performs some different line of planning and handles all the various parts that require that particular operation.

The personnel of our Planning Organization, exclusive of department supervisors and clerks, consists of 86 college graduates, 168 trained men who have come to this organization from our shop departments, or who have had experience in other shops, and 38 men who are neither college graduates nor shop men. The last group of men are mostly those of high school education who have been trained in our line of work.

The requirements of the Planning Engineer on whom the responsibility rests for the successful manufacture of our apparatus are quite extensive. He must first have the ability to plan the manufacture of the apparatus in the most economical manner consistent with the quantity and quality desired and this, of course, cannot successfully be done without a thorough knowledge of the methods and practices necessary in carrying on manufacturing activities along one or more definite lines. He must have a large measure of foresight, thereby reducing to a minimum the difficulties that are bound to occur when the manufacture of a new or changed piece of apparatus is started.

Furthermore, he must make a study of the design of the apparatus under consideration to determine if there are features of it which present manufacturing difficulties either from a tool, assembly, or adjustment standpoint. This part of our work involves a discussion of the manufacturing problems on a new design with the Engineering Organization and the men who handle this work must be able to express themselves in a clear and concise manner to insure that proper consideration is given to the manufacturing suggestions.

It goes without saying that the men who fit best into this organization are those who have had the benefit of an engineering education, preferably specializing on manufacturing methods.

We have, as you will have noted, a large number of planning engineers who have had actual shop experience either with us or in other manufacturing plants, and little or no technical education before working in the shops.

It is noticeable that these men, almost without exception, have realized their handicap due to the lack of a technical education and have either taken advantage of our schools or school work outside.

As stated previously, we have three main sources of supply for the men making up our Planning Organization; first, the Engineering Institutions; second, shop men who have the experience and have to some degree educated themselves in engineering; and third, high school graduates whom we have trained.

Such a combination of trained men makes a strong organization in which the man of superior education and the practical man are mutually helpful to each other in the successful working out of our manufacturing problems.

Irregularities in Loaded Telephone Circuits

By GEORGE CRISSON

SYNOPSIS. The development of long distance telephone transmission has made the question of line irregularities a matter of great importance because of their harmful effect in producing echo currents and causing the repeaters to sing.

The structure of coil-loaded circuits permits the calculation of the probability of obtaining an assigned accuracy of balance between line and network when certain data are known or assumed regarding the accuracy of loading coil inductance and section capacity.

Formulae are given and the results of calculations compared with measurements made on circuits of known accuracy of loading.

INTRODUCTION

THE application of repeaters to telephone circuits in which the speech currents in the two directions of transmission pass through the same electrical path, has caused considerable emphasis to be placed on the matter of making the telephone circuits as free as possible from irregularities. This paper aims to present the theory of the relation between the irregularities in coil loaded lines and the effects resulting therefrom, which have an important bearing upon the operation of two-way telephone repeaters.

The idea of applying the theory of probability to the problem of summing up the effects of many small line irregularities was first suggested in 1912 by Mr. John Mills. The effect upon repeater operation of impedance unbalance had been mathematically analyzed by Dr. G. A. Campbell; and the effect upon impedance of a single irregularity of any type had been investigated by Mr. R. S. Hoyt. Using a probability relationship which was pointed out by Mr. E. C. Molina, Mr. Mills developed a formula which gives the average or probable impedance departure in terms of average or probable irregularities in inductance or capacity, which served at the time of the engineering of the transcontinental line (1913–14) and for some years after.

With the rapid growth of repeatered circuits in cable it became necessary to calculate what fraction of a large number of essentially similar lines would give a definite impedance unbalance at a given frequency. The necessary mathematical work to indicate the conditions for a large group of similar lines was recently carried out independently by Messrs. H. Nyquist and R. S. Hoyt.

The theory which has thus been evolved over a period of years is now presented in a manner which it is hoped will be found relatively simple and useful. Various charts are given which should be of material aid in the application of the theory. There are also given the results of some experiments made on cable circuits in which comparison is made between the impedance departures of the circuits as obtained by direct measurement with the departures as computed from data covering the individual irregularities. These impedance

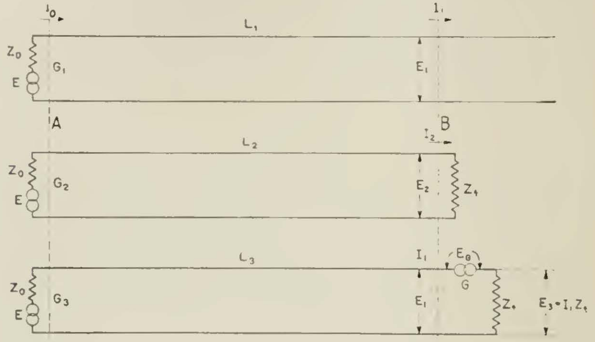


Fig. 1

departures are expressed as "return losses," the meaning of which is explained below. The agreement is shown to be close enough to constitute a good check as to the correctness of the underlying theory.

Magnitude of Reflected Current

In Fig. 1, are shown three regular [1] telephone lines of the same type beginning at a certain point A. The first line L_1 passes through another point B and continues on to infinity. The second line L_2 terminates at B where it is connected to an impedance Z_t which differs from the characteristic impedance Z_o of the three lines, thus constituting an irregular termination. The third line L_3 also terminates at B where it is connected to an impedance Z_t and a generator G of zero impedance whose purpose will be described later. At the sending end A each line is provided with one of three identical generators, G_1, G_2, G_3, having an impedance equal to Z_o the characteristic impedance of the line. The internal voltages of these generators are all equal and represented by E. The generator G_1 impresses a

[1] In this paper the term "regular" implies that a telephone line is free from electrical irregularities.

voltage $E_o = \frac{1}{2} E$ upon the sending end of the line L_1 and causes a current I_o to flow into it. The voltage and current waves are propagated regularly over the line to the point B where they set up a potential difference E_1 between the conductors and cause a current I_1 to flow. E_1 and I_1 are smaller in magnitude and later in phase than E_o and I_o because of the losses and finite velocity of transmission of the line L_1. These quantities have the relation

$$\frac{E_o}{I_o} = \frac{E_1}{I_1} = Z_o \tag{1}$$

since the line is regular.

In the second line L_2 a different set of conditions exists. In this case, the voltage E_2 and the current I_2 produced at B by the generator have the relation

$$\frac{E_2}{I_2} = Z_t. \tag{2}$$

When the e.m.f. of the generator G is zero, the conditions in the third line L_3 are the same as in L_2 but by adjusting the phase and magnitude of the e.m.f. of this generator the current in the terminal impedance Z_t can be made equal to I_1 and the drop across this impedance becomes

$$E_3 = I_1 Z_t. \tag{3}$$

Under these conditions the current I_1 flows at the end of the line L_3 and the potential difference E_1 exists between the conductors at this point. The line L_3 is then in the same condition as the line L_1 between the points A and B. When the waves arrive at B over the line L_3 the generator boosts or depresses the voltage at the terminus of the line by just the amount necessary to cause the terminal apparatus to take the desired current. Then the e.m.f. of the generator G is

$$E_G = E_3 - E_1. \tag{4}$$

Removing the e.m.f. of the generator G makes the conditions in line L_3 identical with the conditions in L_2, but removing this e.m.f. is the same thing as introducing another e.m.f. $-E_G$ in series with the generator which annuls its e.m.f. E_G. This e.m.f. $-E_G$ causes a current I_3 to flow back into the line

$$I_3 = -\frac{E_G}{Z_o + Z_t}. \tag{5}$$

Substituting from equations (1), (3) and (4) above

$$I_3 = \frac{Z_o - Z_t}{Z_o + Z_t} I_1. \tag{6}$$

That is, the effect of connecting an impedance Z_t to the end of a line of characteristic impedance Z_o is to return toward the source a current whose value is $\frac{Z_o-Z_t}{Z_o+Z_t}$ times the current that would exist at the terminus if the line were regularly terminated. The ratio between

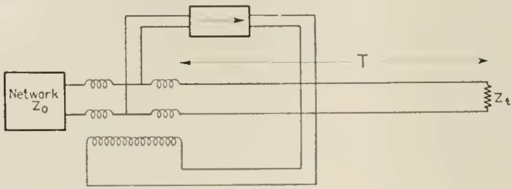


Fig. 2

the reflected and the incident current is known as the "reflection coefficient," the value of which is expressed as follows:

$$r=\frac{I_3}{I_1}=\frac{Z_o-Z_t}{Z_o+Z_t}. \qquad (7)$$

This ratio can also be expressed in transmission units (TU). When expressed in TU this relation will be referred to in this paper as the "transmission loss of the returned current," or, briefly, as the "return loss."

If a condition occurs in a line which causes the impedance at any point to differ from the characteristic impedance it has the same effect as an irregular termination.

Return Loss at a Repeater Due to a Single Irregularity

Fig. 2 shows a No. 21-type repeater connected between a line and a network whose impedance is exactly equal to the characteristic impedance Z_o of the line. If the line is perfectly regular the repeater will be perfectly balanced and the gain can be increased indefinitely without causing the repeater to sing.

Assume now that the line is terminated by some apparatus having an impedance Z_t at a distance from the repeater such that the transmission loss of the intervening line is T TU. If a wave of current having a certain magnitude leaves the repeater, it is reduced in strength by T TU when it reaches the terminus. Of this current, a certain amount is transmitted back toward the repeater, suffering a further loss of T TU on the way; consequently, the relation expressed in TU between the strength of the currents leaving and returning to the repeater, that is, the return loss at the repeater, is given by the equation

$$S = 20 \log_{10} \frac{Z_o + Z_t}{Z_o - Z_t} + 2T. \tag{8}$$

If the gain of the repeater, expressed in TU, is equal to or greater than S the repeater will sing provided the returning current has the correct phase relation to reinforce the original wave. For this reason the term "singing point" has frequently been applied to the quantity S, which is called returned loss in this paper.

If the line is shortened until the impedance Z_t is connected directly to the repeater terminals, the transmission loss T between the repeater and the irregularity is reduced to zero and the return loss becomes

$$S = 20 \log_{10} \frac{Z_o + Z_t}{Z_o - Z_t}. \tag{9}$$

Return Loss of Irregular Lines

In practice, lines are never perfectly regular. Not only is it impracticable to build apparatus which would form a perfectly regular termination for a line, but there are numerous causes of irregularity in the lines themselves, each one of which is capable of reflecting a portion of the waves which traverse the line. These irregularities can be kept smaller than any specified amount if sufficient care is used in building and maintaining the line but they cannot be entirely eliminated; consequently, if a length of actual line is terminated regularly by a network of impedance Z_o, the return loss will be high if the line is carefully built and low if it contains large irregularities. The return loss of such a line, when terminated regularly by a network is a measure of the quality of the line from the standpoint of repeater performance. In measuring the return loss of a line it is necessary that a rather long section of the line be available so as to include all irregularities near enough to have an appreciable effect upon the result. If the section measured is too short, the result will be too high because only a few irregularities will be included.

Calculation of the Return Loss of Coil Loaded Lines

Owing to the facts that the inductance of coil loaded lines is concentrated principally in the loading coils and the capacity is divided into elements of finite size by the loading coils and, further, that the electrical irregularities are due principally to the deviations of the inductance of the coils and the capacity of the sections from their average values for the line, it is possible to calculate by a fairly simple method the value of the return loss of a coil loaded line if the representative values of these deviations and the electrical properties of the line are known or assumed.

Since the return loss depends upon the accidental combination of a large number of unbalance currents there will not be one definite value applying to all circuits, but an application of the theory of probabilities makes it possible to compute what return loss will probably be surpassed by any assigned fraction of a large group of lines having the given deviations.

The method of calculating the return loss of coil loaded lines will now be described. The symbols used in this description and their meanings are given in the following table:

TABLE I

A = Attenuation Factor per Loading Section = Ratio of the Current Leaving a Loading Section to the Current Entering it.

C = Normal Capacity per Loading Section in Farads.

F = Fraction of a Large Group of Lines.

f = Any Frequency for which a Return Loss is to be Found.

$f_c = \frac{1}{\pi\sqrt{LC}}$ = Critical or Cutoff Frequency of the Line.

H_C = Representative[2] Deviation of the Capacity of Loading Sections.

h_C = Deviation of the Capacity of a Particular Loading Section.

H_L = Representative[2] Deviation of the Inductance of Loading Coils.

h_L = Deviation of the Inductance of a Particular Loading Coil.

$H = \sqrt{H_C^2+H_L^2}$ = Representative[2] Combined Deviation.

I_s = Current Entering the Line.

I' = Representative[2] Total In-Phase Returned Current at the Sending End.

I'' = Representative[2] Total Quadrature Returned Current at the Sending End.

I_r = Value of Returned Current which will be Exceeded in a Specified Fraction F of a Large Group of Lines.

i' = Total In-Phase Current at the Sending End of the Line.

i'' = Total Quadrature Current at the Sending End of the Line.

$i_1, i_2, i_3, \text{- - -}\ i_n$ = Currents Returned from the 1, 2, 3, - - - - - and nth Irregularities.

$i_1', i_2', i_3', \text{- - - - -}\ i_n'$ = In-Phase Components of $i_1, i_2, i_3, \text{- - -}\ i_n$

$i_1'', i_2'', i_3'', \text{- - - - -}\ i_n''$ = Quadrature Components of $i_1, i_2, i_3, \text{- - -}\ i_n$

$k = \sqrt{\frac{L}{C}}$ = Nominal Characteristic Impedance of the Line.

L = Normal Inductance of a Loading Coil.

n = Number of Irregularities.

P = Probability Function for the Absolute Value of the Total Returned Current at the Sending End.

p' = Probability Function of the Total In-Phase Returned Current.

R_C = Representative[3] Reflection Coefficient at Capacity Irregularities.
R_L = Representative[3] Reflection Coefficient at Inductance Irregularities.
r_C = Reflection Coefficient at a Capacity Irregularity.
r_L = Reflection Coefficient at an Inductance Irregularity.
$r_1, r_2, r_3, \cdots r_n$ = Reflection Coefficient at the 1, 2, 3, - - - nth Irregularities.
S = Return Loss, Infinite Line.
S_n = Return Loss, Finite Line.
S_A = Attenuation Function.
S_P = Distribution Function.
S_H = Irregularity Function.
S_w = Frequency Function.
T = Transmission Loss in a Finite Line.
$\theta_1, \theta_2, \theta_3, \cdots \theta_n$ = Phase Angles of the Currents at the Sending End Returned by the 1, 2, 3, - - - nth Irregularities.
w = f/f_c.

Reflection at a Coil Irregularity

If a loading coil has too much or too little inductance, the effect is the same as if a small inductance $h_L L$ had been added to or taken away from the coil. The reactance of this increment is $2\pi f L h_L$. The additional reactance has the same effect wherever it may occur in the load but it is somewhat simpler to assume that the increment is introduced at mid-coil. Within the useful range of telephonic frequencies, the mid-coil impedance of a loaded line is given closely by the expression $k\sqrt{1-w^2}$.

In equation (7) $Z_o - Z_t$ corresponds to $2\pi f L h_L$ while $Z_o + Z_t$ is approximately equal to $2k\sqrt{1-w^2}$ when the irregularity is small, consequently:

$$r_L = \frac{\pi f L h_L}{k\sqrt{1-w^2}} \tag{10}$$

and, substituting for f and k their equivalents obtained from relations given in Table I,

$$r_L = h_L \frac{w}{\sqrt{1-w^2}}. \tag{11}$$

Reflection at a Spacing Irregularity

If a loading section has too much or too little capacity, the effect, neglecting conductor resistance, is the same as if a small bridged capacity $h_C C$ were added to or removed from the line. The effect

[3] The "representative" deviation or current is an index of the magnitude of the deviation or current that may be expected in accordance with the laws of the distribution of errors. It corresponds to the root-mean-square error. It must not be confused with the "effective" or r.m.s. value of a particular alternating current. The meaning of the term as used here is more completely explained in the paragraph following equation (24).

is the same for any point in the section, but it is somewhat simpler to assume that the additional capacity is applied at mid-section. The reactance of the added capacity is $\frac{1}{2\pi f h_C C}$ and the mid-section impedance is, closely, $\frac{k}{\sqrt{1-w^2}}$.

When the bridged reactance is large compared with the line impedance, the reflection coefficient r_C is given closely by the equation

$$r_C = \frac{1}{2}\frac{\dfrac{k}{\sqrt{1-w^2}}}{\dfrac{1}{2\pi f h_C C}} \tag{12}$$

from which, substituting the values of f and k as before

$$r_C = h_C \frac{w}{\sqrt{1-w^2}} \tag{13}$$

which is identical in form with equation (11) above.

Approximations Made in Deriving R_L and R_C

The expressions for the mid-coil and mid-section impedances used above in deriving equations (10) and (12) are simple approximations which take no account of the effects of the resistance of the line conductors and loading coils, leakage between conductors or distributed inductance. The errors due to these effects are negligible in the important parts of the frequency range involved in telephone transmission when the types of loading and sizes of conductors now commonly used are considered. The errors due to these causes tend to increase for frequencies which are very low or which approach the cutoff frequency. For accurate calculations relating to very light loading applied to high resistance conductors it would be desirable to take into account the effects of resistance. Because the use of the precise expressions would greatly complicate this discussion and would probably serve no very useful purpose at this time, the approximations given above are used.

Current Returned to the Sending End of the Line

Consider first a line having only one kind of irregularity as, for example, one in which only the loading coils are assumed to vary from their normal values. If a current I_o enters such a line, a current

i_1 is returned to the sending end from the first irregularity (assumed to be very near the sending end)

$$i_1 = r_1 I_o \tag{14}$$

a second current

$$i_2 = A^2 r_2 I_o \tag{15}$$

is returned from the irregularity located at a distance of one loading section away from the sending end, since the current is reduced by the factor A in going to the irregularity and again in returning.

Similarly, a current

$$i_n = A^{2(n-1)} r_n I_o \tag{16}$$

is returned from the nth irregularity.

The first current will return to the sending end with a certain phase angle θ_1 with respect to the initial current, the second with a phase angle θ_2, etc. Each returned current may be resolved into two components, one in phase with the initial current and one in quadrature.

The in-phase components of the currents are then:

$i_1' = I_o r_1 \cos \theta_1$ from the first irregularity. (17)

$i_2' = I_o r_2 A^2 \cos \theta_2$ from the second irregularity. (18)

$i_3' = I_o r_3 A^4 \cos \theta_3$ from the third irregularity. (19)

$i_n' = I_o r_n A^{2(n-1)} \cos \theta_n$ from the nth irregularity. (20)

and the quadrature components are:

$i_1'' = I_o r_1 \sin \theta_1$ from the first irregularity. (21)

$i_2'' = I_o r_2 A^2 \sin \theta_2$ from the second irregularity. (22)

$i_3'' = I_o r_3 A^4 \sin \theta_3$ from the third irregularity. (23)

$i_n'' = I_o r_n A^{2(n-1)} \sin \theta_n$ from the nth irregularity. (24)

Now the deviations of the inductance (and capacity) resemble the errors of measurement discussed in many text books dealing with the precision of measurement, consequently, they can be studied and their effects combined by the same mathematical law.

Examination of measurements of the inductance of large numbers of loading coils and the capacities of the pairs and phantoms in many reels of cable have shown that the most reasonable assumption is that the deviations of inductance and capacity follow the "normal" law of the distribution of errors.

The deviation at each irregularity is not known but it is possible to derive from the measurements of the inductance of large numbers of loading coils (and the capacity of many lengths of cable) representative values for these deviations similar to the "mean error." Because of the way in which the effects of irregularities combine, this *representative deviation* is taken as the square root of the mean of the squares of the deviations (r.m.s. deviation) of the individual coils. If the average deviation of a large group of coils is known, but the individual deviations are not, it may be multiplied by 1.2533 to obtain the representative deviation on the assumption that the deviations follow the normal law of errors.

If then the representative deviation H_L is substituted for the particular deviation h_L in equation (11), we obtain the representative reflection coefficient

$$R_L = H_L \cdot \frac{w}{\sqrt{1-w^2}}. \tag{25}$$

Now in the usual case where no effort is made to select the loading coils and so obtain a special distribution of the deviations the representative deviation and the representative reflection coefficient are the same for each coil. Substituting R_L for r_1, r_2, etc., in equations (17) to (24) each equation gives the representative value, at the sending end of the line, for the current reflected from the corresponding irregularity.

According to the laws for the combination of deviations which are demonstrated in treatises dealing with precision of measurements the representative value of the current due to all the irregularities would be the square root of the sum of the squares of the representative values of the different currents taken separately, consequently the representative in-phase current is

$$I' = I_o R_L \sqrt{(\cos^2\theta_1 + A^4\cos^2\theta_2 + A^8\cos^2\theta_3 + \cdots A^{4(n-1)}\cos^2\theta_n)} \tag{26}$$

and the representative quadrature current is

$$I'' = I_o R_L \sqrt{(\sin^2\theta_1 + A^4\sin^2\theta_2 + A^8\sin^2\theta_3 + \cdots A^{4(n-1)}\sin^2\theta_n)}. \tag{27}$$

By assuming that the representative in-phase and quadrature currents are equal the following steps can be greatly simplified. In view of the varying effects of frequency, distance from the sending end and nature of the irregularity upon the phase relations this appears to be a justifiable assumption, so combining I' and I'' in quadrature,

$$I' = I'' = \sqrt{\frac{I'^2 + I''^2}{2}} = \frac{I_o R_L}{\sqrt{2}} \sqrt{1 + A^4 + A^8 + \cdots A^{4(n-1)}} \tag{28}$$

For a finite number of irregularities, that is a finite line terminated by a perfect network just beyond the nth coil:

$$I' = I'' = \frac{I_0 R_L}{\sqrt{2}} \sqrt{\frac{1-A^{4n}}{1-A^4}} \tag{29}$$

which is obtained by summing up the series of terms under the radical in equation (28).

For an infinitely long line A^{4n} becomes zero since $A<1$ and

$$I'_\infty = I''_\infty = \frac{I_0 R_L}{\sqrt{2}} \sqrt{\frac{1}{1-A^4}}. \tag{30}$$

I' corresponds to the r.m.s. error in the ordinary theory of errors, consequently the probability function for the distribution of the in-phase currents is:

$$p' = \frac{1}{I'\sqrt{2\pi}} e^{-\frac{i'^2}{2I'^2}}. \tag{31}$$

Changing the accents, this equation also applies to the quadrature components.

The probability that the in-phase current lies between two near by values i' and $i'+di'$ is then equal to $p'\,di'$ and the probability that the quadrature component also lies between two values i'' and $i''+di''$ at the same time is $p'di' \times p''di''$. Transferring to polar coordinates,[3] the probability that the total returned current will be between a value $i=\sqrt{i'^2+i''^2}$ and a slightly different value $i+di$ and also have a phase angle between θ and $\theta+d\theta$ is

$$P = \frac{1}{2\pi I'^2} i e^{-\frac{i^2}{2I'^2}}\, di\, d\theta. \tag{32}$$

Integrating with respect to the phase angle θ between O and 2π to find the probability of obtaining a current between i and $i+di$ of any possible phase displacement

$$F = \frac{1}{I'^2}\int_{I_F}^{\infty} i e^{-\frac{i^2}{2I'^2}}\, di. \tag{33}$$

Integrating between I_F and infinity gives the probability that the total returned current will exceed the value I_F.

$$F = e^{-\frac{I_F^2}{2I'^2}}. \tag{34}$$

[3] For a more complete description of this operation, see "Advanced Calculus," by E. B. Wilson, page 390 et seq.

In a large number of lines, F is the fraction of the whole group which will have a return current in excess of I_F.

From the definition of the transmission unit the return loss of the line expressed in TU, is given by the expression

$$S = 20 \log_{10} \frac{I_o}{I_F} = -20 \log_{10} \frac{I_F}{I_o} \tag{35}$$

from which

$$I_F^2 = I_o^2 10^{-\frac{S}{10}}. \tag{36}$$

Substituting in (34)

$$F = e^{-\frac{I_o^2}{2I'^2} 10^{-\frac{S}{10}}} \tag{37}$$

Taking logarithms to the base e and transposing

$$10^{-\frac{S}{10}} = -\frac{2I'^2}{I_o^2} \log_e F. \tag{38}$$

Taking logarithms to the base 10

$$S = 10 \log_{10} \left[\frac{I_o^2}{2I'^2 \log_e \frac{1}{F}} \right]. \tag{39}$$

Substituting the value of I'_∞ from equation (30) for I'

$$S = 10 \log_{10} \left[\frac{1-A^4}{R_L^2} \times \frac{1}{\log_e \frac{1}{F}} \right] \tag{40}$$

and the value of R_L from equation (25)

$$S = 10 \log_{10} \left[\frac{1}{H_L^2} \times \frac{1-w^2}{w^2} \times (1-A^4) \times \frac{1}{\log_e \frac{1}{F}} \right]. \tag{41}$$

By a similar process of reasoning it is evident that if the line contains capacity deviations only, the return loss is given by this same expression with H_C substituted for H_L and if both types of irregularity occur the representative deviation is

$$H = \sqrt{H_L^2 + H_C^2}$$

when H_C includes the effect of spacing irregularities as well as capacity deviations in the cable. The foregoing expression can, for convenience, be put in the form

$$S = S_H + S_w + S_F - S_A \tag{42}$$

in which each term depends upon only one independent variable and in which the symbols have the following meanings:

$$S_H = \text{Irregularity function} = 20 \log_{10} \frac{1}{H} \tag{43}$$

$$S_w = \text{Frequency function} = 20 \log_{10} \frac{\sqrt{1-w^2}}{w} \tag{44}$$

$$S_F = \text{Distribution function} = 10 \log_{10} \frac{1}{\log_e \frac{1}{F}} \tag{45}$$

$$S_A = \text{Attenuation function} = 10 \log_{10} \frac{1}{1-A^4} \tag{46}$$

Meaning of Equation (42)

To understand more clearly the meaning of equation (42) imagine that a large number of circuits of the same type and gauge are to be built in accordance with the same specifications so that the representative (r.m.s.) deviation including all causes has the same value H for each circuit. Further, imagine that the value of S has been calculated by formula (42) using a particular frequency f and a convenient fraction F. It is to be expected that when the circuits have been built and their return losses measured at the given frequency f the fraction F of the whole group will have return losses lower than S and the rest will have higher return losses.

In discussing expected results it is sometimes preferable to state the fraction $1-F$ of the circuits whose return losses will be greater than the assigned value rather than the fraction F whose return losses will be lower. This is done in Figs. 9 to 14 described below.

Location of the First Irregularity

In equations (14), (15) and (16) and all the equations which depend upon them it was assumed that the first irregularity occurs at the sending end of the line. Two other assumptions are equally plausible and might under some circumstances be preferable. These are that the first irregularity occurs (a) at one-half section from the end or (b) at a full section. In the first case (a) the current returned to the sending end from each irregularity will be reduced by the factor A and in the second (b) by the factor A^2, that is the return loss given by equation (42) should be increased by (a) the amount of the transmission loss in one loading section or (b) twice the amount of the transmission loss in one loading section respectively.

Return Losses of Short Lines

When a line is short and regularly terminated the returned current will be somewhat less than if it extends to infinity with irregularities and consequently the return loss will be higher. From equations (29) and (30), the returned current is lowered in the ratio $\frac{I'}{I'_\infty} = \sqrt{1-A^{4n}}$ by limiting the line to n sections; consequently

$$S_n = S + (S_n - S) = S + 10 \log_{10} \frac{1}{1-A^{4n}} \tag{47}$$

in which

$$S_n - S = 10 \log_{10} \frac{1}{1-A^{4n}} \tag{48}$$

is the increase in return loss.

Since the transmission loss in n sections of the line is

$$T = 20 \log_{10} \frac{1}{A^n} \tag{49}$$

it is easily seen that the increase of return loss can be expressed as a function of this loss. Transposing (49) and substituting in (48)

$$S_n - S = 10 \log_{10} \frac{1}{1 - \left[\frac{1}{\log^{-1}_{10} \frac{T}{20}}\right]^4} \tag{50}$$

Charts

The process of computing return losses can be greatly shortened by using the graphs of equations (43), (44), (45), (46), and (50) to obtain the values of the various functions. The accompanying Figs. 3 to 8, inclusive, have been prepared to illustrate these graphs and for use in rough calculations.

S_H may be obtained from any table or chart giving the relation between TU and current ratio by using H like a current ratio. Fig. 3 is a chart drawn especially for this purpose. For values of H lying between 0.1 and 0.01 look up a point on the curve corresponding to $10H$ and add 20 TU to the corresponding value of S_H, for values of H lying between 0.01 and 0.001 look up a point corresponding to $100H$ and add 10 TU to the value of S_H, and so forth.

Figs. 4, 5, 6, and 7 are curves giving the relations between the functions S_w, S_F and S_A, respectively, and the quantities upon which


IRREGULARITY FUNCTION - TU
SH = 20 Log10 1/H
SH - TU
20
15
10
5
0
.1
.15
.2
.3
.4
.5
.6
.7
.8
.9
1.0
H

Fig. 3

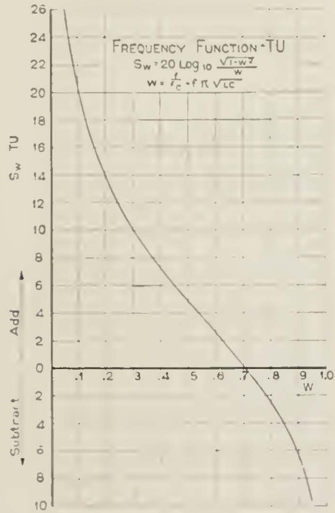

FREQUENCY FUNCTION - TU
Sw = 20 Log10 √(1-w²)/w
w = f/fc = f π √LC
Sw TU
Add
Subtract
26
24
22
20
18
16
14
12
10
8
6
4
2
0
2
4
6
8
10
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0
W

Fig. 4

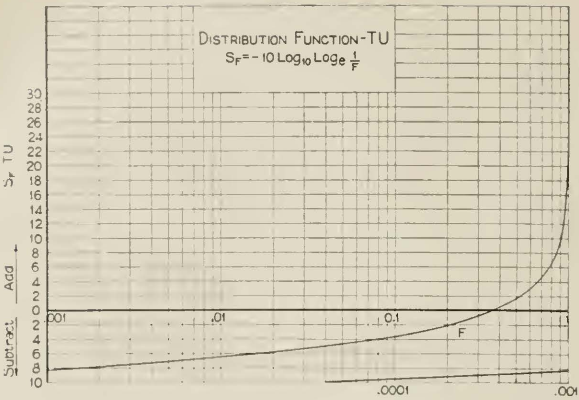


Fig. 5

Attenuation Function—TU
In terms of loss per loading section

$$S_A = 10 \log_{10} \frac{1}{1-A^4},$$

A = Attenuation factor per loading section,

$$L = 20 \log_{10} \frac{1}{A} = \text{loss per loading section in TU}$$

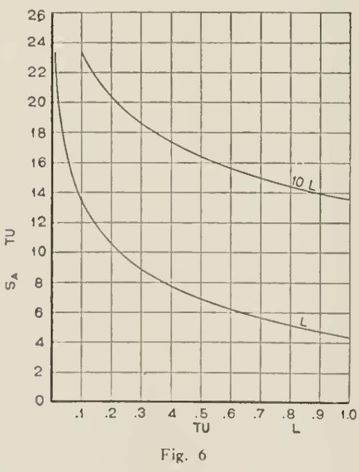


Fig. 6

each depends plotted from equations (44), (45) and (46). These are all positive except as indicated by the word "Subtract" on the diagrams.

A simple method for extending the curve of Fig. 5 is as follows: (a) choose a point on the curve within 3 TU of the lower end, (b) subtract about 3 TU (accurately, $10 \log_{10} 2$) from the value of S_F for this point, and (c) square the value of F for this point. The results obtained for (b) and (c) are the coordinates of another point on the extension of the curve.

Fig. 6 gives the relation between S_A and the transmission loss per loading section. On account of the wide use of 6,000 ft. spacing the curves of Fig. 7 are plotted to give the relation between S_A and the transmission loss per mile for 6,000 ft. spacing which is usually a more convenient arrangement.

Fig. 8 gives the amount, $S_n - S$, by which the return loss of a regularly terminated line of finite length (n sections) is greater than that of an infinite line as a function of the transmission loss of the finite line. This was calculated by formula (50).

Calculation of Return Loss

The process of finding the return loss by means of the curves is as follows:

(1) Determine the value of H_L, the representative deviation of the loading coils, and H_C, the representative deviation of the capacity of the loading sections. These depend upon the variations allowed in the specifications for loading coils and cable and upon the care with which the line is built. Calculate $H = \sqrt{H_L^2 + H_C^2}$, the representative combined deviation of the section. Look up the number of TU corresponding to H in any suitable table or chart, such as Fig. 3, to find S_H.

(2) Assume the frequency, f, to be considered. Calculate $w = \frac{f}{f_c}$ and look up the corresponding value of S_w on Fig. 4.

(3) Assume a value of F and look up the corresponding value of S_F on Fig. 5.

(4) Look up the value of S_A on Fig. 7, corresponding to the transmission loss per mile of the circuit at the frequency f if the coils are spaced 6,000 feet (1.136 miles) apart, or calculate the loss per section and look up S_A on Fig. 6, if some other spacing is used.

(5) Calculate $S = S_H + S_w + S_F - S_A$.

Attenuation Function—TU

In terms of loss per mile of the circuit length of loading section 6000 ft.

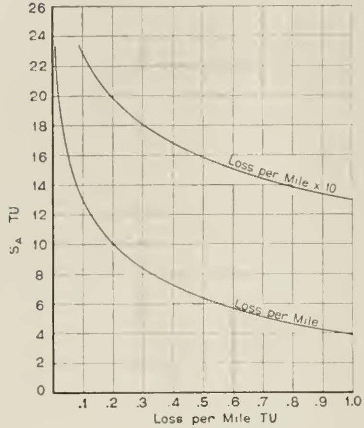


Fig. 7

Increase of the return loss when the line is limited to *n* sections

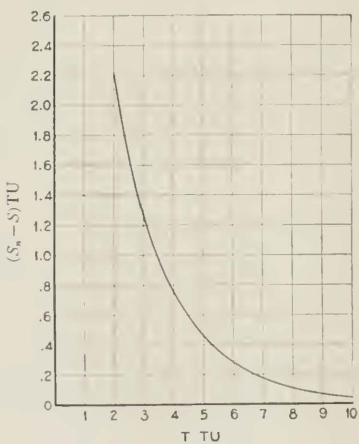


Fig. 8

(6) If the return loss of a finite length of line is desired determine the transmission loss of this length and look up the corresponding value of $S_n - S$ on Fig. 8. Add this amount to the value of S found in paragraph (5).

Example

As an example to illustrate the application of these methods let us calculate a return loss at 1,000 cycles for No. 19-H-174-63 [4] side circuits such that 90 per cent. of the circuits may be expected to have a higher value and only 10 per cent. to fall below it. The necessary data are given in Table II, below.

(1) $H = \sqrt{0.0062^2 + 0.0129^2 + 0.0045^2} = 0.0150.$

Fig. 3 gives 36.5 TU as the corresponding value of S_h.

(2) At 1,000 cycles $w = \dfrac{1000}{2810} = 0.356$

Fig. 4 gives 8.4 TU as the corresponding value of S_w.

(3) Since 90 per cent. of the finished lines are to have return losses greater than S and 10 per cent. less $F = 0.1$ and Fig. 5 gives -3.7 TU as the corresponding value of S_F.

(4) The transmission loss per mile is 0.271. Since the coils are spaced 6,000 feet apart, Fig. 7 gives 8.7 TU as the value of S_A. This same value would be obtained less directly by calculating the loss per loading section, $0.271 \times \dfrac{6000}{5280} = 0.311$ and using Fig. 6. The latter method is used when the spacing is different from 6,000 feet.

(5) Using equation (42)

$$S = S_H + S_w + S_F - S_A = 36.5 + 8.4 - 3.7 - 8.7 = 32.5 \text{ TU}.$$

This will be found to agree with the 90 per cent. point on the smooth curve plotted in Fig. 10 which is described below.

(6) In case it is desired find the return loss of a length of this line having a transmission loss of, for example, 6 TU instead of the return loss of the infinite line. Fig. 8 gives $S_n - S = 0.3$ from which

$$S_n = 32.5 + 0.3 = 32.8 \text{ TU}.$$

Determination of Tolerable Deviations

To determine the deviations which correspond to an assigned value of the return loss find values of S_w, S_F and S_A as in paragraphs (2),

[4] In accordance with the practices of the Bell System, this notation indicates a phantom group of No. 19 B. & S. conductors in a cable with loading coils spaced 6,000 feet apart, the side circuit coils having 174 millihenrys inductance and the phantom coils 63 millihenrys.

(3) and (4) above and substitute in formula (42) to find the value of S_H. This with a table or chart of TU and current ratio gives the value of H. Limits can then be imposed on the loading coil inductances and section capacities that will insure that the representative deviation will not exceed the value H so found.

Comparison of Different Types of Circuits

These formulae are useful in comparing the return losses to be expected in various types of circuits which are built with the same accuracy in the matters of coil inductance and section capacity. In such cases it is merely necessary to calculate the quantity $S_w - S_A$ for each circuit and take the difference.

Example

As an example compare the No. 19-H-174-63 side circuits worked out above with No. 16-H-44-S [5] circuits at 1,000 cycles. Since the deviations and the fraction F are the same only S_w and S_A need be considered. For the No. 16-gauge circuit $f_c = 5560$ and the loss in TU per mile is 0.236. From these figures:

Gauge of Line	No. 19	No. 16
$w = \frac{1000}{f_c}$	0.356	0.18
S_w TU	8.4	14.8
S_A TU	8.7	9.4
$S_w - S_A$ TU	−0.3	5.4

These figures show that the return loss of the No. 16-H-44-S circuits should be higher than that of the No. 19-H-174-63 side circuits and the difference to be expected is $5.4 - (-0.3) = 5.7$ TU.

When the circuits to be compared have the same cutoff frequency the process of comparison is even simpler since the quantity S_w is then the same in each case. S_A is determined for each circuit as in paragraph (4) above. The difference between the two values of S_A is the difference between the return losses.

Example

As an example compare the No. 19-H-174-63 side circuits with No. 16-H-174-63 side circuits. In this case the cutoff frequencies are the same so w and S_w are the same. It is then only necessary to compare S_A. The loss per mile of the No. 16-gauge circuit is 0.161

[5] This notation indicates a side circuit of No. 16 B. & S. conductors in a cable loaded with 44 millihenry coils spaced 6,000 feet apart.

TU at 1,000 cycles from which $S_A = 11$ TU. In equation (12) S_A is negative hence the No. 19-gauge will have a higher return loss than the No. 16-gauge circuits and the expected difference is $11 - 8.7 = 2.3$ TU.

Comparison of Calculated and Measured Return Losses

In order to test the methods of calculation described above a series of measurements of return loss at 500, 1000 and 2000 cycles were made on a group of loaded side and phantom circuits in a cable using a No. 2-A unbalance measuring set.

The representative inductance deviations were found by analyzing the inductance measurements on a large group of loading coils similar to those used in the cable. The representative capacity deviations, not including the spacing irregularity were found by analyzing the shop measurements on a number of reels of the cable. This gave representative figures for reel lengths which were divided by $\sqrt{12}$ (in accordance with the laws of probability since this cable had 12 reel lengths in a loading section) to obtain the representative capacity deviations due to the cable for the loading sections. The spacing deviations were separately determined from the measured distances between the loading points.

The data used in the calculation were as follows:

TABLE II

		Sides	Phantoms
Representative inductance deviation		0.0062*	0.0061*
Representative capacity deviation		0.0129*	0.0138*
Representative spacing deviation		0.0045*	0.0045*
Combined representative deviation, H		0.0150*	0.0158*
Cutoff frequency *fc* (cycles sec.)		2810	3727
Transmission loss TU per mile	500 cycles	0.265	0.271
	1000 cycles	0.274	0.279
	2000 cycles	0.317	0.296

The smooth curves of Figs. 9 to 11, inclusive, were calculated from the data in Table II using the methods described above. The abscissas are the percentages of a large group of circuits which may be expected to have return losses greater than the values given by the ordinates. This percentage is equal to 100 $(1-F)$. The points plotted on the

* The figures are "fractional" deviations. Percentage deviations which are sometmes used are 100 times as large. Care should be taken to avoid errors caused by failure to divide percentage deviations by 100 before finding the value of F_H.

Return loss of No. 19-H-174-63 sides exceeded by various percentages of circuits at 500 cycles

Smooth curve—theoretical

⊙ 46-H-174-63 sides Pittsburgh to Ligonier

⊡ 12-H-174-63 sides Ligonier to Pittsburgh

△ 52-H-174-106 sides Pittsburgh to Ligonier

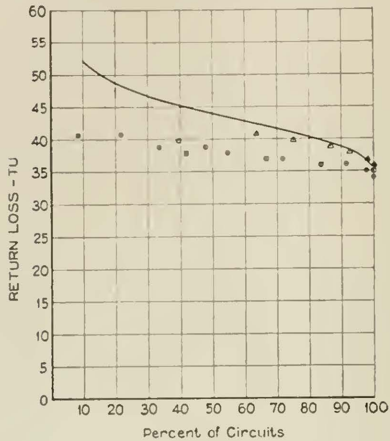


Fig. 9

Return loss of No. 19-H-174-63 sides exceeded by various percentages of circuits at 1000 Cycles

Smooth curve theoretical

⊙ 46-H-174-63 sides Pittsburgh to Ligonier

⊡ 12-H-174-63 sides Ligonier to Pittsburgh

△ 52-H-174-106 sides Pittsburgh to Ligonier

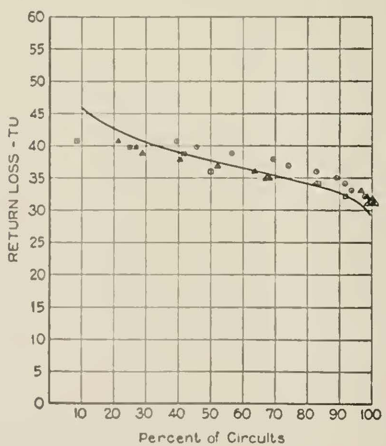


Fig. 10

Return loss of No. 19-H-174-63 sides exceeded by various percentages of circuits at 2000 cycles

Smooth curve—theoretical

⊙ 16-H-174-63 sides Pittsburgh to Ligonier

⊡ 12-H-174-63 sides Ligonier to Pittsburgh

△ 52-H-174-106 sides Pittsburgh to Ligonier

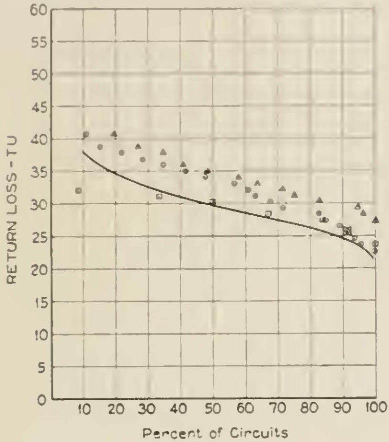


Fig. 11

Return loss of No. 19-H-174-63 phantoms exceeded by various percentages of circuits at 500 cycles

Smooth curve—theoretical

⊙ 25-H-174-63 phantoms Pittsburgh to Ligonier

⊡ 21-H-174-63 phantoms Ligonier to Pittsburgh

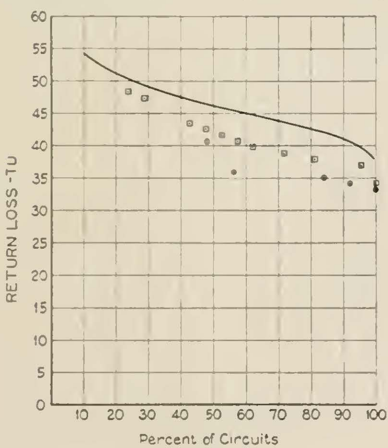


Fig. 12

Return loss of No. 19-H-174-63 phantoms exceeded by various percentages of circuits at 1000 cycles

Smooth curve—theoretical
⊙ 25-H-174-63 phantoms Pittsburgh to Ligonier
⊡ 21-H-174-63 phantoms Ligonier to Pittsburgh

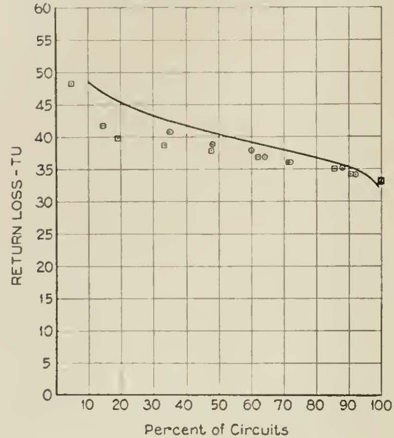


Fig. 13

Return loss of No. 19-H 174 63 phantoms exceeded by various percentages of circuits at 2000 cycles

Smooth curve—theoretical
⊙ 25-H-174-63 phantoms Pittsburgh to Ligonier
⊡ 21-H-174-63 phantoms Ligonier to Pittsburgh

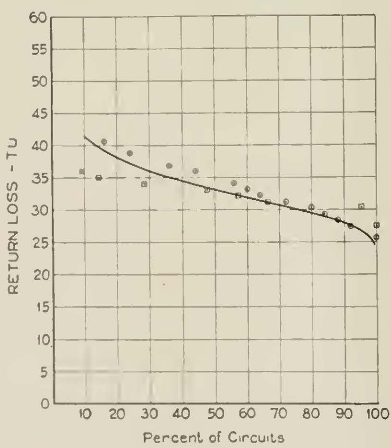


Fig. 14

curve sheets give the measured values of return loss found in the groups of circuits listed in the explanatory notes on the drawings.

In general, it will be observed that there is a fair agreement between the theoretical curves and the measured return losses especially at 1000 and 2000 cycles.

Due to the limited range of the measuring apparatus, readings of return losses greater than 40.7 TU were not made except in the case of the Ligonier to Pittsburgh phantoms shown on Figs. 12, 13 and 14, when a special arrangement was available to extend the range to 47.3 TU. For this reason points representing observed return losses above these limits are not available which causes the observed values for 500 cycles in Figs. 9 and 12 to appear somewhat low at first sight.

Where the highest point in a given set of data represents many circuits as in the cases represented by the small triangles and circles in Fig. 9 this point probably gives closely the return loss corresponding to the percentage of circuits it indicates but the points for higher return losses are not available. When the highest point represents only one or two circuits as in the case represented by the square in Fig. 9, it is likely that the actual return loss is higher than the point indicates.

It should also be noted that above 40 TU the actual impedance of the line and its characteristic impedance differ by less than 2 per cent. so that very small departures of the network from the true characteristic impedance of the line would tend to make the observed return loss low.

Conclusion

It is believed that the procedure described in this paper offers a reliable method for determining the probability of attaining a particular value of return loss at any assigned frequency when a circuit is built with definite limitations on inductance and capacity deviations so that the representative deviations are known.

The Sounds of Speech

By IRVING B. CRANDALL

NOTE: As professor of vocal physiology, Alexander Graham Bell did pioneer research in "devising methods of exhibiting the vibrations of sounds optically." In 1873, he became familiar with the phonautograph, developed by Scott and Koenig in 1859, and with the manometric capsule, developed by Koenig in 1862. Greatly impressed by the success of these instruments "to reproduce to the eye those details of sound vibration that produce in our ears the sensation we term timbre, or quality of sound" Bell used an improved form of the phonautograph having a stylus of wood about a foot long. He obtained "large and very beautiful tracings of the vibrations of the air of vowel sounds" upon a smoked glass.

In describing his early attempts to improve the methods and apparatus for making speech waves visible and to interpret wave form, Bell wrote:

"I then sang the same vowels, in the same way, into the mouth-piece of the manometric capsule, and compared the tracings of the phonautograph with the flame-undulations visible in the mirror. The shapes of the vibrations obtained in the two ways were not exactly identical, and I came to the conclusion that the phonautograph would require considerable modification to be adapted to my purpose. The membrane was loaded by being attached to a long lever, and the bristle, too, at the end of the lever, seemed to have a definite rate of vibration of its own. These facts led me to imagine that the true form of vibration characteristic of the sounds of speech had been distorted in the phonautograph by the instrumentalities employed. I therefore made many experiments to improve the construction of the instrument. I constructed, at home, quite a number of different forms of phonautographs, using membranes of different diameters and thicknesses, and of different materials, and changing the shape of the attached lever and bristle."

Struck by the likeness of the phonautograph and the mechanism of the human ear, Bell conceived the idea of making an instrument modeled after the pattern of the ear, thinking it would probably produce more accurate tracings of speech vibrations. In 1874, he consulted a distinguished aurist, Dr. Clarence Blake of Boston, who suggested that instead of trying to make an instrument modeled after the human ear, the human ear itself be used. Dr. Blake prepared a specimen containing the membrane of tympanum with two bones attached, the malleus and incus. The other bone, the stapes, was removed and a stylus of wheat straw about one inch long was substituted. A sort of speaking tube was arranged to take the place of the outer ear. "When a person sang or spoke to this ear, I was delighted to observe the vibrations of all the parts and the style of hay vibrated with such amplitude as to enable me to obtain tracings of the vibrations on smoked glass."

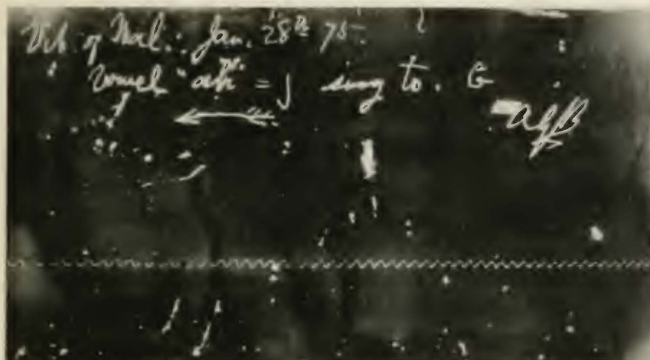
In the accompanying paper, Dr. I. B. Crandall describes modern methods whereby with the most refined apparatus, highly accurate speech wave forms have been produced. The analysis and interpretation of both vowel and consonant sounds made possible by these records, are the realization of an objective sought by Bell a half century ago.

This article is the result of an extended study of 160 graphical records of vowel and consonant sounds, of which a few are reproduced in the present publication. One hundred and four of these records are of vowel sounds and formed the basis of the "Dynamical Study of the Vowel Sounds," by I. B. Crandall and C. F. Sacia which was published in this Journal in April, 1924. The purpose of the present article is to describe all of the records in sufficient detail, including in one discussion the outstanding characteristics of vowel, semi-vowel and consonant sounds; it is hoped shortly to supplement this with a reproduction of a larger group of records from the complete collection.—*Editor.*

CONTENTS

Introduction

TO the layman speech is a matter of course, but to the student of science, or of language "the amazing phenomenon of articulate speech comes home . . . as a kind of commonplace miracle." [1] Hence we have inquiries into the nature of speech from many points of view, beginning with fundamentals based on physiology and acoustic



Speech record made by Bell in 1875

science and leading to important applications in communication engineering, phonetics and vocal music.

The scientific study of speech sounds began with Helmholtz, who also made a fundamental study of hearing. Helmholtz had the advantage, in approaching these problems, of a knowledge of physiology as well as a mastery of theoretical physics. With this equipment and such simple laboratory apparatus as he created, he did his great work on speech and hearing of which we have the record (in English translation) under the title of "Sensations of Tone." [2] Today, with

[1] Greenough & Kittredge, "Words and Their Ways," N. Y., 1901.

[2] "The Sensations of Tone as a Physiological Basis for the Study of Music." Translated from the Fourth German Edition by A. J. Ellis: Fourth English Edition, London, 1912.

immeasurably superior physical apparatus, and with more specialized theoretical equipment, the individual investigator usually approaches one problem at a time, the problem and the method being selected according to the technique with which he is familiar. The work of D. C. Miller on sound and sound analysis [3] represents the beginning of modern physical research on speech sounds. In medical science some attention has been given to the mechanism of speech [4] and the psychologists are responsible for an enormous literature on voice control and the perception of speech and tones.[5] The work of Scripture [6] represents the beginning of a science of experimental phonetics, and in the closely related field of philology there is a rapidly growing interest in the physical characteristics of speech sounds.[7]

In this large field of investigation the physicist finds a real opportunity in providing means for the study and measurement of speech sounds, and a real responsibility in broadening the extent and improving the accuracy of such quantitative data as are obtained.

The results obtained from such physical investigations have practical as well as scientific value, and we observe that in a large laboratory concerned entirely with the development of electrical communication considerable effort has been devoted to research on speech and acoustic apparatus.[8] It has recently been felt that the wave

[3] "The Science of Musical Sounds," New York, 1916. This contains a bibliography of 90 special references, some 12 of which relate specifically to speech.

[4] "A Contribution to the Mechanism of Articulate Speech," by S. W. Carruthers. Edin. Med. Jour. VIII (New Series) (1900) pp. 236, 332, 426.

[5] "The Psychology of Sound," by Henry J. Watt (Cambridge, England, 1917), contains a bibliography of 159 references. The work of C. E. Seashore is noteworthy in this field.

[6] "Researches in Experimental Phonetics." Publication No. 44, Carnegie Institution, Washington, 1906.

[7] "The Physical Characteristics of Speech Sound," by Mark H. Liddell. Bulletin No. 16, Purdue University Engineering Experiment Station.

[8] See following papers, from the Research Laboratories of the American Telephone and Telegraph Co. and Western Electric Co., Inc.:

(a) H. D. Arnold and I. B. Crandall: The Thermophone as a Precision Source of Sound: Phys. Rev. 10, (1917), p. 22.
(b) E. C. Wente: Condenser Transmitter for Measurement of Sound Intensity: Phys. Rev. 10 (1917), p. 39.
(c) I. B. Crandall: The Air Damped Vibrating System: Phys. Rev. 11 (1918), p. 449.
(d) I. B. Crandall: The Composition of Speech: Phys. Rev. 10 (1917), p. 74.
(e) R. L. Wegel: Theory of Telephone Receivers: J. A. I. E. E. 40 (1921).
(f) E. C. Wente: Sensitivity and Precision of the Electrostatic Transmitter: Phys. Rev. 19 (1922), p. 498.
(g) I. B. Crandall and D. Mackenzie: Analysis of the Energy Distribution in Speech: Phys. Rev. 19 (1922), p. 221.
(h) H. Fletcher: The Nature of Speech and its Interpretation: J. Franklin Inst. 193 (1922), p. 729.
(i) J. Q. Stewart: An Electrical Analogue of the Vocal Organs: Nature, Sept. 2, 1922.

forms of the speech sounds required more precise determination, and indeed research in the art of telephony has emphasized this need. The graphical records of speech sounds, which form a supplement to the present paper, are contributions to this study.

I

Note on the Characteristic Frequencies of Speech

Speech is, in itself, a sound wave a succession of condensations and rarefactions in the air. For the purposes of this study we are not primarily concerned with the mechanism of production, nor with the processes of perception of speech, though it may be necessary to digress to inquiries of this kind, in their bearing on certain characteristics of speech. We are interested primarily in what can be learned from the records of the speech vibrations themselves.

Speech sounds are complex, that is, they are composites of simple sounds, each component having a particular frequency, amplitude, phase and duration. Considering speech in the mass, we find its energy distributed among frequencies from 75 to above 5,000 cycles with the larger part of this energy contained in the region below 1,000 cycles. This distribution is shown approximately in Fig. 1 taken from reference (8g); the limitation on these data being that the measuring apparatus was not sufficiently sensitive to measure the speech energy associated with frequencies higher than 5,000 cycles. Inasmuch as the energy of speech resides largely in the vowel sounds, the curve in Fig. 1 can also be taken as applying to the average distribution in the vowel sounds. The energy distribution diagram is of fundamental importance in the physical study of speech sounds; it reveals at once the frequencies of large energy content which are characteristic. For each vowel sound, there is a distinctive energy frequency diagram.

The consonant sounds present a difficult problem because of the small amount of energy associated with them. Most of our knowledge of the consonant sounds is qualitative: for example Fletcher (reference 8h) who studied the nature of speech by the method of testing articulation when different frequency ranges are eliminated shows that for two fricative or sibilant consonants *s* and *z*, there are essential frequency components which lie above 5,000 cycles. The characteristic frequencies of the consonant sounds are usually only part of the whole story; these sounds are richer in transients, and clearly less periodic in their nature than the vowel sounds. And in between the two broad classes of consonant and vowel sounds there is a group of semi-vowel sounds (*r*, *l*, *m*, *n*, *ng*) closely related to the vowel group, and yielding readily a determination of their "characteristic frequencies."

There are two physical theories of vowel production; and these two theories suggest different methods of analyzing the vowel sounds into components of simpler nature. These two points of view we

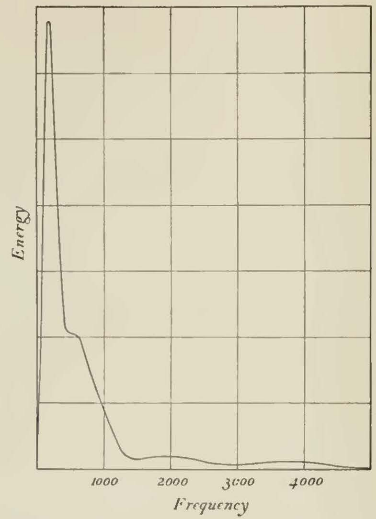


Fig. 1—Energy distribution; composite curve for male and female voices

shall briefly consider along historical lines. We are indebted to Helmholtz for the greatest single contribution to the study of the vowels, in that he gave a complete diagram of the characteristic frequencies of the vowels (ref. 2, pp. 103–109), which was based on his celebrated experiments in analysis and synthesis by means of the Helmholtz resonators. But in connection with his scheme of characteristic frequencies he took up the theory of Wheatstone (1837) that these frequencies are true harmonic components of the cord tones, which were reenforced by resonance in the oral cavities. Some later physicists have followed this so-called *harmonic or steady state theory* of the vowel sounds, notably Miller (reference 3, pp. 239–243) who made a very careful study of the whole matter. According to this theory the obvious procedure is to apply the classical Fourier analysis to determine the characteristic components of the vowel sounds.

Turning now to the other (and earlier) view, the so-called "Inharmonic Theory" of Willis (1829) later developed by Hermann and rather recently by Scripture (ref. (6)) we are invited to believe that the "characteristic frequencies" of the vowel sounds are the natural vibrations or *transients* in the oral cavities, when excited impulsively by the (more or less) periodic puffs of air from the glottis. According to this theory no harmonic relations need obtain between the characteristic frequencies of the vowels and the fundamental or cord tone accompanying them; and the classical Fourier analysis is not considered applicable in resolving the vowel sound into simpler components. According to this "inharmonic" or "transient" theory we must treat the natural vibrations of the oral cavities as damped vibrations and find the frequencies and damping constants of their components, as best we can from the record of the complete sound vibration.

In favor of the Helmholtz or "Harmonic" theory we have the careful studies by Helmholtz and his successors of the relations between the cord or fundamental tone, its harmonics as reenforced by the oral cavities or other resonators, and the observed characteristic frequencies of the vowel sounds. The oral cavities constitute a vibrating system of two or three degrees of freedom, the theory of which has been fully developed by Rayleigh and others, and it is to be expected that, with the speaking mechanism in normal adjustment the vowel qualities can be well accounted for by postulating harmonic forced vibrations in these cavities. This expectation has been realized in the numerous successful attempts which have been made to produce vowels artificially by using a harmonic series of tones, and reenforcing certain harmonics by suitable resonators. Miller's experiments with organ pipes (ref. 3, pp. 246–250), in which he successfully reproduced certain vowel sounds, are well known.

The Willis-Hermann theory has also suggested much notable experimental work. Scripture (ref. 6, p. 114) constructed a "vowel-organ" in which a reed pipe was used to excite the natural vibrations in resonators designed to imitate the conditions in the oral cavities, and attained some success in reproducing vowels. More recently J. Q. Stewart (ref. 8) has produced an "Electrical Analogue" of the vocal organs with which remarkable results in reproducing vowel sounds and even some of the consonant sounds have been obtained. In this electrical arrangement transients excited by an interrupter in oscillatory circuits take the place of the transient vibrations of the oral cavities. Finally Paget (reference (9a) below) has constructed a whole series of double resonators which may be excited by blowing air into them through an "artificial larynx," and from which he has obtained all of the vowel sounds. As the result of this work he has given a very complete chart of the characteristic frequencies of the vowels and he has been led to the conclusion that there are *two* characteristic frequencies or regions of resonance for each vowel sound.

From the standpoint of practical acoustics both theories have contributed to progress, and it seems that the experimental physicist would not be justified in partiality to either view. Speech is a variable phenomenon; the cord tones are not always stable; in speaking and in singing there are allowable variations in duration, intensity and frequency of the component tones without essential change in the characteristics of the vowel sounds. Given accurate records of the speech sounds as normally pronounced by a number of speakers, we should expect to arrive at nearly the same characteristic frequencies whichever mode of analysis we adopt. As pointed out by J. Q. Stewart (Ref. 8i) Rayleigh has stated (Sound, Vol. II, p. 473) that the disagreement between the Helmholtz-Miller, or steady state theory of vowels, and the Willis-Hermann-Scripture, or transient theory is only apparent; to quote Stewart, "The disagreement concerns methods rather than facts. Which viewpoint should be adopted is thus a matter of convenience in a given case. When the transmission of speech over telephone circuits is in question, for example, the steady state theory often possesses obvious mathematical advantages. On the other hand, the quantitative data relating to the physical nature of vowels which are given in D. C. Miller's well-known book "The Science of Musical Sounds" expressed as they are in terms of the steady state theory are less compact and definite than the data of Table I (Stewart's paper) which are expressed in terms of the transient theory. The general agreement between the two sets of data is, of course, obvious."

In studying the behavior of vibrating systems from the theoretical standpoint, there is a tendency to emphasize the intimate relations that exist between transient and steady state phenomena. Both depend only on the driving forces and the constants of the system,

[9](a) Sir R. A. S. Paget: "The Production of Artificial Vowel Sounds." Proc. Roy. Soc. A102, Mar. 1, 1923, p. 752.

[9](b) A second memoir: "The Nature and Artificial Production of Consonant Sounds." Proc. Roy. Soc. A 106, Aug. 1, 1924, p. 150, to which reference will be made in more detail later.

Other papers by Paget include: Nature, Jan. 6, 1923, "Nature and Reproduction of Speech Sounds." Electrician, Apr. 11, 1924. The Same Title. Proc. Land. Phys. Soc 36 pt 3, Apr. 15, 1924, p. 213: Discussion on Loud Speakers.

hence "the solution for transient oscillations of the system is reduced to formulae which are functionally the same as those for steady state oscillations" (reference 10; see also reference 11). But before leaving this discussion of speech characteristics it should be noted that the essence of the matter lies not so much in reconciling the two theories of the vowel sounds as in ascertaining what motions really take place in the oral cavities, and in the air near the vocal cords. Though the process of harmonic analysis is to be applied to the records of the vowel sounds, we must recognize its limitations, and not necessarily infer steady state conditions. Indeed the most casual inspection of the records shows a certain *lack* of periodicity in the phenomena recorded; and it is hardly to be expected that all the phenomena can be satisfactorily summed up on the basis of the harmonic theory.

II

The Recording Apparatus[12]

In providing means for accurately recording sound waves, use has been made of three devices recently developed in this Laboratory and we believe that by properly connecting these together we have obtained a recording instrument which is superior in accuracy and power to any heretofore used. These three devices were each nearly free from distortion, and such residual distortions as could not be eliminated were so controlled that they practically offset one another over a wide range of frequencies.

The first element in the recording set is the condenser transmitter, which has been thoroughly investigated by Wente (refs. 8b, 8c, 8f); its frequency characteristics, in both amplitude and phase are shown in Fig. 2. The particular transmitter used was of recent design and had been carefully standardized and calibrated especially for this work.

The condenser transmitter was connected to the input terminals of a seven-stage amplifier as shown in the large diagram of Fig. 5 which gives the details of the electrical circuit, including the third

[10] J. R. Carson: Phys. Rev. X, 1917, p. 217, "On a General Expansion Theorem for the Transient Oscillations of a Connected System."

[11] T. C. Fry, Phys. Rev. XIV, 1919, p. 117. "The Solution of Circuit Problems."

[12] Thanks are due to Messrs. C. F. Sacia and C. J. Beck for the skill and care with which they assembled and calibrated the recording apparatus, and made the complete set of records. The writer is also under obligation to Mr. Sacia for aid in choosing the sounds to be recorded, and systematizing the collection; Mr. Sacia also developed and applied the photomechanical method of analyzing records, the results of which are given in Figs. 13 and 14 of this paper.

element, a special oscillograph, which was connected to the output terminals of the amplifier. The first six tubes, in cascade, provided a voltage amplification of about 40,000; the last eight tubes, in parallel, constituted a "current transformer" working into the low impedance of the oscillograph vibrator, with a small resistance in series. The coupling between the stages, and between amplifier and terminal apparatus,

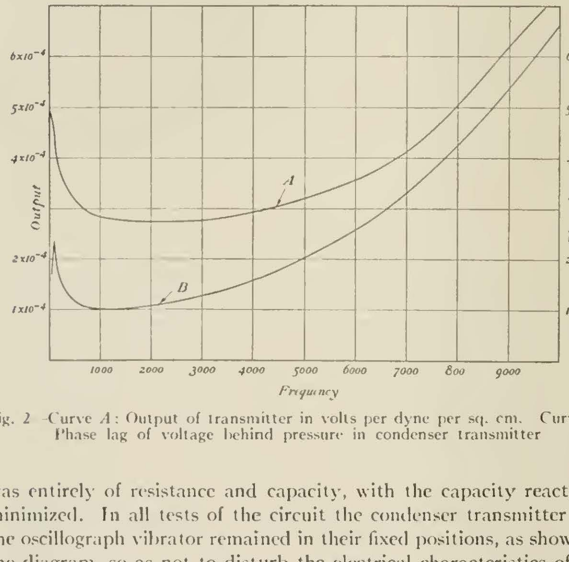


Fig. 2—Curve *A*: Output of transmitter in volts per dyne per sq. cm. Curve *B*: Phase lag of voltage behind pressure in condenser transmitter

was entirely of resistance and capacity, with the capacity reactance minimized. In all tests of the circuit the condenser transmitter and the oscillograph vibrator remained in their fixed positions, as shown in the diagram, so as not to disturb the electrical characteristics of the circuit. The frequency characteristics of the amplifier in amplitude and phase are shown in Fig. 3. In measuring the amplitude characteristic a small electromotive force was introduced in series with the transmitter, in the input mesh; and in measuring the phase lead of the output as a function of frequency use was made of the Alternating Current Potentiometer of Wente (Jour. A. I. E. E. Dec. 1921) the other details of procedure being as usual.

The characteristics of the oscillograph vibrator are shown in Fig. 4. This vibrator was specially constructed, with small mass, high tension and damping; when the requisite dynamical characteristics were once obtained, its calibration presented no great difficulty.

In combining the transmitter, the amplifier and the oscillograph to form the complete recording apparatus there were two primary requirements; first, the set as a whole should be free from frequency distortion

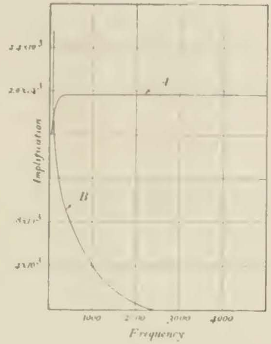


Fig. 3—Curve *A*: Amplitude frequency characteristic of amplifier. Curve *B*: Phase lead of output, vs. frequency of voltage input to amplifier

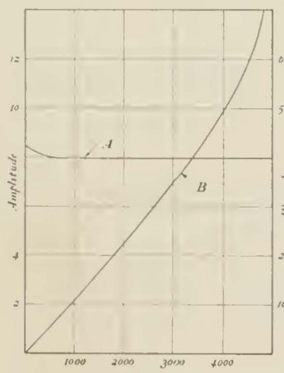


Fig. 4—Curve *A*: Amplitude frequency characteristic of oscillograph. Curve *B*: Phase lag of amplitude behind current in oscillograph

in both amplitude and phase, and second, the output of the set as a whole should be a linear function of the input within the working energy range at each frequency. The first of these conditions is in general the harder to fulfil. Frequency-amplitude distortion has been practically eliminated as we have seen from each of the three essential parts of this apparatus; and although it was found impracticable to make each part of the apparatus free from frequency distortion in phase, it was possible to give the complete set good frequency characteristics in both amplitude and phase as will be explained.

In a vibrating system of one degree of freedom when we wish to avoid frequency distortion in amplitude, we usually adjust the resonant

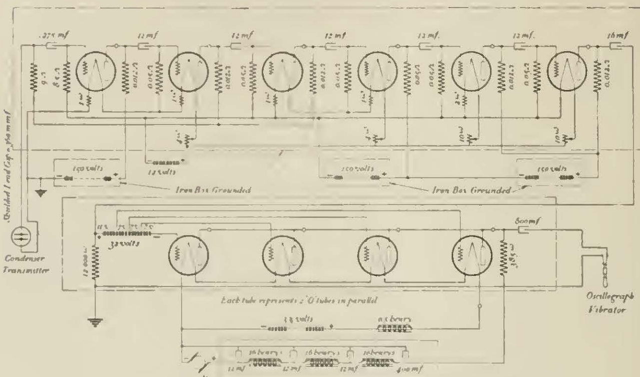


Fig. 5—General diagram of recording apparatus showing circuit details

frequency so that it is above the range of frequencies within which we desire to work; in addition, it is desirable in most cases to make the damping of the system large. With these adjustments made it is found that there is a phase lag between amplitude and driving force which rises with frequency and reaches a maximum above the resonant frequency, and it is possible to make this phase lag nearly proportional to the frequency over the range of frequencies within which it is desired to work.

It is well known that if equal driving forces produce equal amplitudes at all frequencies, and if the phase lag of the amplitude with respect to the driving force is proportional to frequency, then a driving force of complex wave form is reproduced without distortion of wave form in the vibrating system. These conditions held very well over the desired range of frequencies in the oscillograph vibrator, as shown in Fig. 1. In the case of the condenser transmitter, however, there were departures from these conditions in the frequency interval from zero to 500 cycles for which allowance had to be made.

In the amplifier the effect of capacity reactance was nearly eliminated. Owing to the small remaining capacity reactance there was a phase lead of amplifier current with respect to driving force which was applied to offset the excessive phase lag in the condenser transmitter at the low frequencies. The particular adjustment of amplifier finally arrived at represented the best compromise, considering the difficulty

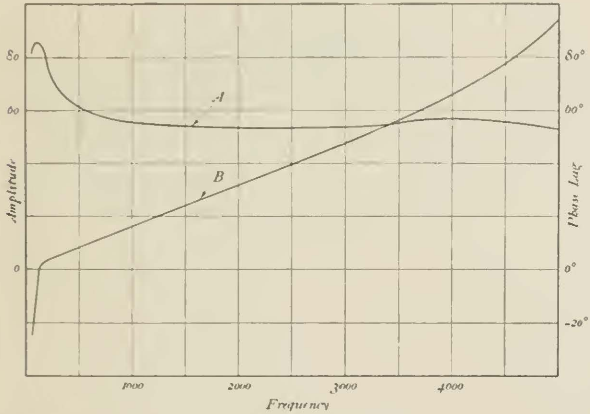


Fig. 6—Overall frequency characteristics of amplitude and phase of the recording system. Curve A: Oscillographic amplitude per unit of pressure on transmitter diaphragm. Curve B: Phase lag of oscillographic amplitude behind pressure on diaphragm

encountered with the transmitter characteristics. With this compromise made there was an unavoidable phase lead in the whole apparatus for frequencies below 125 cycles, but this was not serious as most of the speech energy is in higher frequencies. After all final adjustments were made the overall frequency characteristics of amplitude and phase were as shown in Fig. 6. Thus ultimately there was obtained a system with practically uniform amplitude characteristic from 500 to 5,000 cycles, without serious departure from this level for frequencies from 50 to 500 cycles; and with phase lag nearly a linear function of frequency from 125 to 5,000 cycles, after passing through a period of lead in the narrow interval from 50 to 215 cycles.

Consider now the second requirement which the recording system had to meet: namely, that the output of the system should be a linear function of the input within the working energy range at each frequency. Thorough investigation of the condenser transmitter had shown that this instrument met this second requirement very well; it was only necessary to test the remainder of the system. Fig. 7 gives

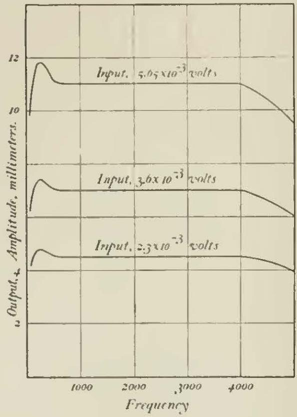


Fig. 7—Amplitude frequency characteristics of circuit-oscillograph at different energy levels

the results of these tests, the voltages introduced in series with the transmitter at the input being maintained at different constant levels, while the frequency was varied. An inspection of the data shows that this requirement was very accurately fulfilled, by the whole electrical system.

Returning now to the overall characteristics of the apparatus, it was thought advisable to test the calibrations in amplitude and phase lag by comparing the computed and the observed distortion when a square-topped acoustic wave was impressed on the apparatus. The steep sides and the flat tops of these waves can be reproduced without distortion only if the apparatus possesses first class characteristics, both in amplitude and phase lag, and the test was a severe one. As would be expected from the calibration curves of Fig. 6 there was a certain amount of distortion in recording this wave, and the square-topped wave, with its very large fundamental component, made this distortion appear much worse than would an ordinary speech wave.

Fig. 8 illustrates the apparatus used to produce the acoustic square-topped wave. An electrode resembling the back plate of the condenser transmitter was mounted in front of the transmitter diaphragm. Between this electrode and the diaphragm was applied a high potential which was made alternately positive and negative by a commutator.

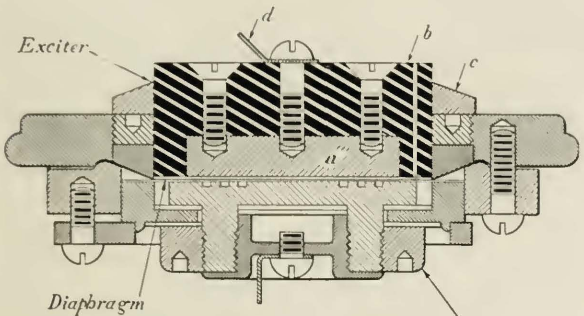


Fig. 8—Condenser transmitter coupled with square-topped-wave exciter

EXCITER PARTS
a. Steel Electrode 0.006 inch from Diaphragm. *b*. Micarta Insulation. *c*. Supporting Ring. *d*. Electrode Terminal.

By this arrangement the desired positive and negative pressures were produced on the diaphragm. The distance between the auxiliary electrode and the transmitter diaphragm was about .006 inch. This electrostatic coupling was found to be sufficiently close to give a suitable deflection of the transmitter diaphragm, while the stiffness and damping of the air film did not alter the dynamical characteristics of the transmitter.

Fig. 9 is an oscillogram showing the wave form recorded by the apparatus when acoustic square-topped waves of frequencies 84, 153 and 306 cycles per second are impressed on the transmitter. Timing waves of frequencies 75, 150 and 300 are also shown. Analysing the original wave by the Fourier method, and allowing for the distortion in amplitude and phase of each component frequency, a computation has been made of the wave form in the output in the case of the square-topped waves of 84 and 153 frequency. The results are shown in Fig. 10.

The Fourier series representing the 84-cycle wave contained 30 terms, the component frequencies being odd multiples of 84 up to a limit of 4,956 cycles; for the series representing the 153-cycle wave 17 terms were used covering the range from 153 to 5,049 cycles. The agreement between calculated and observed output waves would have been more exact, particularly at the corners of the wave shapes, if calibrations

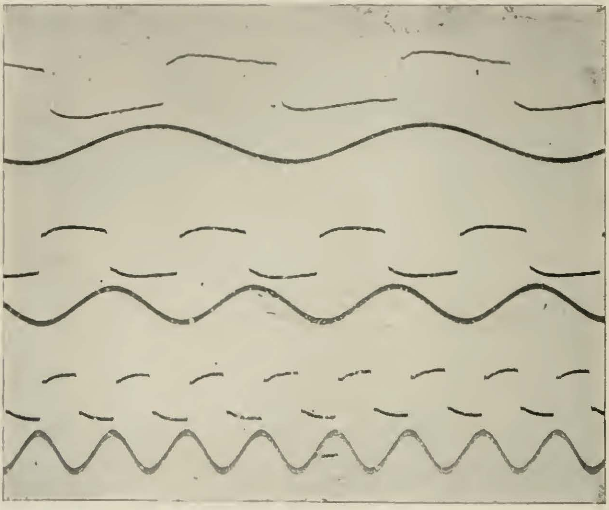


Fig. 9—Oscillogram of square-topped acoustic waves as recorded by the apparatus

and calculations had been carried to frequencies considerably above 5,000. As it was, the performance was considered good; it indicated that the uncorrected records of speech waves as taken were sufficiently accurate for most purposes, while if harmonic analysis of the records was planned accurate results could be obtained over the range from 80 to 5,000 cycles, if the correction factors determined by the calibration were applied.

In this description of the recording apparatus the emphasis has been placed on the dynamical characteristics of the apparatus and its calibration, but some of its other working features may briefly be mentioned. The apparatus was sufficiently powerful to record sounds spoken in an ordinary tone of voice, with the speaker's mouth about three inches from the transmitter. A key was pressed by the speaker just before the sound was spoken, this releasing a shutter placed before a rotating film drum on which the record from the oscillograph vibrator was traced. The film drum was some 52 inches in circumference, and there was mounted on it a length of Eastman super-speed

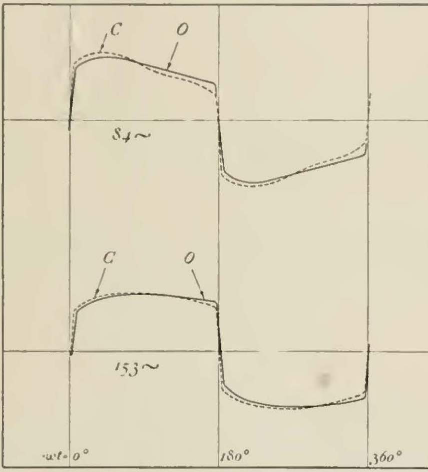


Fig. 10—Calculated and observed wave forms, as recorded by the apparatus

film with which records could be made at a peripheral speed of about 20 feet per second. Thus each hundredth of a second corresponds to two inches or more in the time scale on the film. Besides opening the shutter, the key released a mechanism which swung the oscillograph vibrator through an arc during the progress of the record, thus tracing a helical record on the film. By this means records up to 200 inches in length, or for nearly one second of duration were taken. The average length of the wave trains recorded was less than 0.5 second; thus it was possible to graph the pressure wave of the whole speech sound from beginning to end. Immediately following the recording of the speech sound a timing wave of 1,000 cycle alternating current, taken from a standard oscillator, was recorded on the film at one side of the speech record, without disturbing the speed adjustment of the rotating drum. Thus the time scale was accurately determined for each record.

Especial care was taken with the optical system to insure fine definition and strong illumination of the spot on the film and the films were developed for maximum contrast. As a result, the records were sufficiently clear to permit their reproduction by the line-engraving

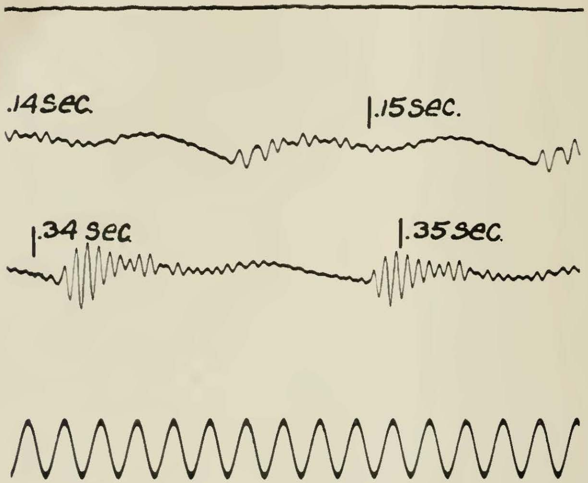


Fig. 11—Section of original record showing timing wave

process. Each of the plates shown in this paper is made up of overlapping sections from the original record, each faithfully reproduced, and the whole arranged to give the complete record within the limits of one page. A section of one of the original records as taken is shown in the figure above.

III

Classification of the Records

In selecting and classifying the vowel sounds for record, use has been made, with slight alteration, of the phonetic arrangement adopted by Fletcher (ref. 8 h). This arrangement of the vowel sounds is illustrated in the diagram of Fig. 12. In this diagram eleven standard "pure-vowel" sounds from *oo* to long *e* are arranged according to the conventional "triangle" and two related vowel sounds *ar* and *er* are interpolated in their proper places. A group of eight records was made of each of these thirteen vowel sounds, four in each group by

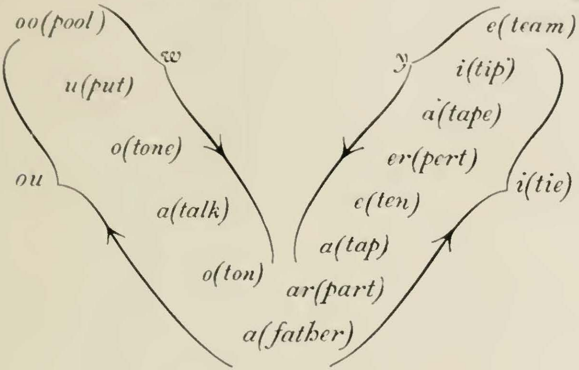


Fig. 12—Classification of vowel sounds

male voices, and the other four by female voices. Each of these records, Plates 1 to 104 (Groups I to XIII), represents the vowel sound as spoken naturally, and continuously recorded from beginning to end.

No attempt was made to record the vowels *w*, *y*, *ou* and long *i*. These usually have transitional characteristics which are sufficiently indicated by the arrows in the diagram. The first two of these, when followed by vowels, and the last two, in nearly all cases, fall into the class of diphthongs.

Following the groups of records of the "pure-vowel" sounds of the diagram it was originally planned to make a group of records of the semi-vowels *l*, *m*, *n*, *ng*, and *r*, recorded in connection with certain vowels. It seemed best however to present records for the sounds *ar* and *er* in connection with the standard vowel sounds as noted above (*ar*, *er*, Groups VII, X) and only these records of the sound *r* were taken. The four remaining sounds were arbitrarily divided into two groups because of the number of records made, and the first of these (Group XIV) contains records of *l* and *ng*. These were made by two male speakers, using the syllables *loo*, *lee*, *la* and *ngoo*, *ngee*, *nga*.

Group XV is devoted to the semivowels *n* and *m*, each recorded with the three vowel sounds *oo*, long *e* and *a*, by the two male speakers, as in the preceding group. Groups XIV and XV are intimately related, and as will appear the four semi-vowel sounds are closely related to the vowel diagram.

When this study was planned, it was thought that the apparatus would be particularly adapted to recording vowel sounds and no great hopes were entertained of applying it to definitive investigation of the consonant sounds. As the work progressed however, it was found that some of the characteristics of the consonant sounds could be recorded and the program was enlarged to include the records of Groups XIV to XVII inclusive. Each of the records of a consonant and vowel combination can be compared with the corresponding record, by the same speaker, of the pure vowel alone in one of the earlier groups, and certain conclusions as to the nature of the consonant sound can be formed.

Group XVI includes records of the six stop (or "hard") consonants *b*, *p*; *d*, *t*; *g*, *k*; followed by two transitional consonants *dth* (as in *then*) *th* (as in *thin*); each associated with the vowel *a*, and recorded by the two male speakers. The natural arrangement is in pairs, the related voiced and unvoiced variations being grouped together.

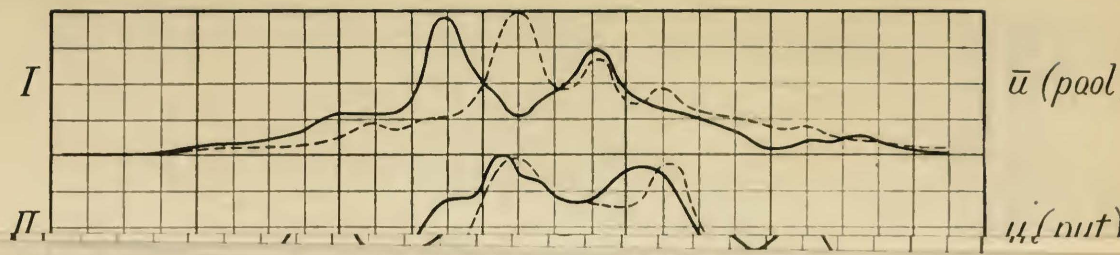
The last Group (XVII) includes records of eight fricative ("soft" or "sibilant") consonants paired in the same way. These are *v*, *f*; *j*, *ch*; *z*, *s*; *zh* (azure), *sh*; each associated with *a* and recorded by the two male speakers.

The following table lists in groups all the records made. As it is not practicable to engrave and print with this article the whole set of

TABLE I

Complete List of Speech Records

Group			Plates
I	*oo* as in *pool*,	by Eight Speakers.	1– 8
II	*u* as in *put*,	by Eight Speakers.	9– 16
III	*o* as in *tone*,	by Eight Speakers.	17– 24
IV	*a* as in *talk*,	by Eight Speakers.	25– 32
V	*o* as in *ton*,	by Eight Speakers.	33– 40
VI	*a* as in *father*,	by Eight Speakers.	41– 48
VII	*ar* as in *part*,	by Eight Speakers.	49– 56
VIII	*a* as in *tap*,	by Eight Speakers.	57– 64
IX	*e* as in *ten*,	by Eight Speakers.	65– 72
X	*er* as in *pert*,	by Eight Speakers.	73– 80
XI	*a* as in *tape*,	by Eight Speakers.	81– 88
XII	*i* as in *tip*,	by Eight Speakers.	89– 96
XIII	*e* as in *team*,	by Eight Speakers.	97–104
XIV	Semi Vowels *l*, *ng*	by two male speakers	105–116
XV	Semi Vowels *n*, *m*	by two male speakers	117–128
XVI	Six Stop Consonants, transitional *dth*, *th*; by two male speakers		129–140
XVII	Eight Fricative Consonants, by two male speakers		145–164


I
ū (paal)
II
u (nut)

160 records, a selection has been made of some 13 typical examples which illustrate characteristic consonant and vowel wave forms. These are listed in table II and their properties are described in detail in the following sections. It may not be amiss to summarize here the basis on which these particular records were chosen for publication.

TABLE II

List of Records Shown in This Paper

Record No.	Plate No.	Title	Speaker
143	9	*u* as in *put*	MA
192	10	*o* as in *ton*	FD
139	11	*a* as in *father*	MA
151	19	*ar* as in *part*	MA
148	89	*i* as in *tip*	MA
234	108	*lee*	MB
238	110	*la*	MB
229	124	*moo*	MB
286	136	*ta*	MB
289	138	*ga*	MB
272	151	*cha*	MA
293	158	*za*	MB
294	160	*sa*	MB

The most important sound (*a*, as in father) is represented in 7 of these records, which include six instances of its combination with other sounds. The record of *ar* (Plate 19) which was chosen is the most characteristic and interesting one of its group. The other vowel records (Plates 9, 10, 89) are sufficiently scattered about the vowel triangle to give an idea of the variation in the high frequency characteristics which is to be an important subject of discussion later. One record of a female voice (Plate 10) is probably sufficient to show the distinctive fundamental, about an octave higher, characteristic of such records. Plate 108 was chosen to show the resemblance between l and e, which establishes a natural transition between the vowel and semi-vowel sounds. From plates 108, 110 and 124 a good idea of the relative amplitudes of vowel and semi-vowel sounds can be obtained; a similar observation holds in the comparison of the vowel and consonant sounds of Plates 136, 138, 151, 158 and 160. Plates 136 and 138 show two extended transients of moderate frequency, the latter in connection with a voiced consonant (*hard g*); Plate 151 is similar to 136 but the vowel following the consonant is less suddenly produced. The pair, Plates 158 and 160, show the voiced and unvoiced hiss (z and s respectively) a sound of very high frequency, which is the limiting case of this type of consonant.

The plates reproduced with this paper are reduced slightly (15 or 20 per cent) in scale, as compared with the original records, to bring them within the page height of the Journal.

In producing this system of records we believe that we have covered the speech sounds as fully as we are justified in doing with the present recording apparatus. In the case of each vowel the combined data from the eight records constitute a sufficient basis for the most thorough harmonic analyses that can be made and they should yield accurate results for the characteristic vowel frequencies. In analysing these records small corrections are of course necessary on account of the slightly imperfect frequency characteristics of the apparatus, but these corrections can be taken without difficulty from the calibration curves.

The amplitude scale in these records is arbitrary in each case. This is for the reason that, owing to the widely different conditions of voice control among the different speakers, the recording apparatus had to be adjusted to different levels of sensitiveness for each record in order to obtain the requisite maximum oscillation of from 1 to 2 centimeters. No attempt has been made to compare the absolute amplitudes from one record to another on account of these intensity variations. The emphasis has been placed rather on obtaining in each record a good well-defined wave which could be enlarged if necessary.

Notwithstanding the fact that for frequencies above 5,000 cycles the apparatus was not nearly as good as for frequencies within the calibration range from 75 to 5,000 cycles, the records obtained of some of the consonant sounds are of considerable practical value. It is felt however, that the present apparatus has been used nearly to the limit of its possibilities and that devices other than the usual oscillograph vibrator offer more promise in any further investigation of the consonant sounds. It is planned later to issue a more complete set of these records as a supplement to the present paper in order to make the collection available to those especially interested.

IV

Statistical Study and Harmonic Analysis of the Vowel Sounds

A detailed inspection of the records taken, and particularly of the records of the vowel groups shows that much labor would be required to analyze these records throughout their length, according to the usual methods of harmonic analysis. In nearly every case it would be impossible to obtain the mean energy distribution in a given record, allowing for variations from cycle to cycle of the fundamental, by choosing from each record only a few such cycles as representative and analyzing these.[1] If, for example, only 10 cycles were taken at selected intervals from each of the 104 vowel records shown there would be required over one thousand such analyses, and to be of value these analyses should include components of frequency from 100 to 5,000 cycles. For this reason a mechanical method of analysis has been applied to determine from the records the average frequency spectra of each of the vowel and semi-vowel sounds.

First let us consider the vowel records in a simpler and more general way. Considerable information has been obtained by inspection, using such simple apparatus as a pair of compasses and a rule in connection with the time scale on the records. The time scale greatly facilitates the process; it is in most cases possible to count the number of cycles of any one prominent component occurring in an interval of .01 second, and by doing this in various parts of the record, to arrive at a rough average frequency for the component in question.

In the case of the low frequency components (the fundamental and the lower characteristic frequency) the procedure was to make this examination at 3 points; one near the start, one near the middle, and one near the end of each record. In this way the most significant changes in pitch and wave form during the course of the record can be brought to light, and some of the individual characteristics of the speaker revealed. A statistical compilation of these results serves to show certain "normal" characteristics of pitch variation, and permit the detection of a certain amount of "personal bias" of the individual speaker in his departure therefrom. In the examination of the low frequency characteristics a note was made as to the harmonic relation between the fundamental and the lower characteristic frequency; of the amplitude of the lower characteristic frequency as being greater or less than the amplitude of the fundamental; and of the behavior of the amplitude of the lower characteristic, during the cycle of the fundamental. The amplitude of the low frequency characteristic is either substantially constant during the cycle or falls away as a transient vibration.

The high frequency components are clearly shown in the records, but it is more difficult to determine their exact frequencies, and practically impossible to relate them harmonically to the fundamental. These oscillations were counted in from four to eight locations in each

[1] It is practicable, however, to obtain valuable data as to the formation of the vowel sounds by analyzing separately the successive cycles at the beginning of a typical vowel record. A study of this kind, based on these records, is being carried out by Messrs. N. R. French and W. Koenig of the American Telephone and Telegraph Company.

record, and a maximum and minimum figure determined for the frequency wherever possible. The behavior of the amplitude of the high frequency component during the cycle was noted, and a rough estimate made of its magnitude. Practically all the vowel records show frequencies above 2500 cycles and the amplitudes in some cases are large. In only two records out of 101 was the high frequency component too small in amplitude to give a frequency determination. These high frequency components may or may not be characteristic of the given sound; this question is more fully dealt with later.

To complete the examination of each record its duration was noted, and this time was divided into three intervals: (1) a building up period in which the oscillations rise from zero to an amplitude which shows all the components clearly; (2) a middle period in which the general amplitude remains nearly constant, but in which some variations in the amplitudes and phases of the component frequencies usually take place; and (3) a period of decay in which the components disappear and the oscillation gradually loses its characteristic wave form.

The procedure may be illustrated by its application to the first record for which the following data were recorded:

Plate No. 1, *oo* as in *pool*. Speaker MA. (Male).

Time to build up, .05 sec.; Middle period, .20 sec.; Period of decay, .06 sec.; Total Duration .31 sec.

Fundamental: 102 at start, rises to 108 in middle, rises to 120 at end. Pitch Variation normal. (See explanation below).

Low Frequency Characteristic: 400 at start, 430 at middle, 440 at end. Amplitude greater than that of fundamental. Approximately, a fourth harmonic of fundamental, but amplitude variation during the cycle suggests a transient.

High Frequency Component: Minimum, 3300 cycles. Maximum, 3600 cycles. Noticeable throughout; amplitude variation suggests a transient.

No other frequencies.

This routine was applied to each of the 101 vowel records and a general summary made of the results, giving approximate values of the vowel characteristics which forecasted the more accurate results obtained later from the mechanical harmonic analysis.

The simplest phenomena to summarize are the general characteristics of the individual speakers. These are based on the mean performance of each in speaking the thirteen vowel sounds, and will be useful in the discussion to follow; they are shown in Table III, below:

TABLE III

Speakers' Characteristics

Male Speakers	Mean Fundamental Pitch at Start, Middle and End	Mean Pitch	Mean Duration of Records
MA—low pitched	97–105–111 (normal)	104	.275 sec.
MB—low pitched	112–115–112 (biased)	113	.222 biased toward short records
MC—high pitched	124–131–134 (normal)	130	.235 (biased toward short records)
MD—high pitched	134–148–175 (normal)	152	.305
	Mean for male Speakers	125	.259 sec.
Female Speakers			
FA—low pitched	224–241–209 (normal)	224	.290 sec.
FB—low pitched	256–251–194 (biased)	234	.373 biased toward long records
FC—medium	233–255–244 (normal)	244	.320
FD—high pitched	271–274–279 (biased)	275	.348 (biased toward long records)
	Mean for female speakers	244	.333 sec.
	Mean duration		.296 sec.

These records were made without constraint imposed on the speaker, except that he had to start and stop within an interval of about one second, and was requested to repeat the sound several times at what he judged to be constant loudness. The resulting variation in performance may therefore be of some interest.

Of 52 men's records the vowel sounds 35 records showed a "normal" effect of progressive *rise in pitch* during the course of the record. (The mode is taken as the normal effect, and follows the mean very closely.) In 6 records out of 13, speaker MB showed an individual or biased effect of slight fall in pitch toward the end. The women's records show greater variation, 24 records out of 52 showing a "normal" effect of a *rise in pitch, followed by falling pitch,* during the course of the record. The individual bias of speaker FB toward progressive fall in pitch was shown in 7 records; that of FD toward progressive rise in 4 records.

The relative constancy in fundamental pitch shown by speaker MB is best exemplified in Plate No. 58. Speaker FD made 3 records of constant pitch: Nos. 21, 40 and 48. Other records of constant pitch are Nos. 19 and 99, both by MC.

In duration, the bias of speaker MB towards short records was shown in 6 records which fell short by .08 sec. or more of the mean for the particular sound considered; that of MC also in 6 records according to the same test. Speaker FB produced 5 records, and speaker FD, 2 records too long by the same amount.

Consider now the general properties of the spoken vowel sound, as deduced from these records. First there is a period of rapid growth in amplitude, lasting about 0.01 second, during which all components are quickly produced, and rise nearly to maximum amplitude; second the middle period, the characteristics of which have been noted, lasting about 0.165 second, followed by the period of gradual decay lasting about 0.09 second, bringing the total length to approximately 0.295 second. There is a tendency to short duration among the "short" vowels (eg. short *o*, *e*, *i*) and a tendency to longer records among the broader sounds, as might be expected.

The behavior of the fundamental frequency (or "cord tone") during the course of the record will follow normal or individual characteristics as has been described.

The low frequency characteristic appears early, usually before the fourth cycle (for men) or before the seventh (for women) and normally is in harmonic relation with the fundamental. In the eleven pure vowel sounds (omitting the *ar* and *er* groups) this point was examined at 264 locations in 88 records with the result that the harmonic relation obtained in at least 214 cases. On the other hand the normal behavior of the amplitude of the low frequency characteristic suggests the decay of a transient oscillation during each fundamental cycle—this effect being noticeable in at least 64 of the 88 pure vowel records. This transient effect was also noticeable in 13 of the 16 records of *ar* and *er*, where the harmonic effect was not so noticeable. The appearance of the transient effect depends to some extent on the relative frequencies of the fundamental and the characteristic; where the fundamental period is short, (as often in the case of the women's records) there is not sufficient time for decay of the characteristic tone before it receives a new impetus in the next cycle of the fundamental.

As noted above, all the records contain high frequency vibrations which are of such amplitude that they suggest characteristic frequencies. A general mean of these frequencies would be in the neighborhood of 3200 cycles, and in the case of two records by speaker FC (Group I and Group XIII) the frequency rises to about 5000 cycles. Recalling the usual classification of the vowel sounds into two groups—(1) those of "single" resonance, placed on the left leg of the triangle, (Fig. 12) and (2) those "double" resonance placed on the right leg of the triangle—there are some differences in the behavior of the high frequency components which can be related to these broad classes.

TABLE IV

Statistical Data From 104 Records of Vowel Sounds

Sound	Duration				Mean Fundamental Frequency		Mean Low Characteristic Frequency		Scattered Low Freq.		Mean High Characteristic Frequency		Scattered High Freq.	
	Start	Middle	Decay	Total	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
I *oo* (pool)	061	164	126	351	140	270	411	581	750 (1)	1200 (1)			3700 (4)	3412 (4)
II *u* (put)	057	115	077	249	138	250	457	691	988 (4)	1100 (3)			3637 (4)	4250 (4)
III *o* (tone)	053	139	133	325	116	237	520	729	830 (3)	1112 (4)			3475 (1)	3700 (4)
IV *a* (talk)	034	191	065	.290	112	243	722	801	950 (2)	1150 (2)			3612 (4)	4075 (4)
V *o* (ton)	046	.179	051	280	118	253	654	854	1100 (4)	1188 (4)			3212 (4)	3353 (3)
VI *a* (father	029	199	078	306	113	234	955	1036	1150 (2)	1425 (2)			3683 (1)	1200 (3)
VII *ar* (part)				345	110	231	630	701	917 (3)	1012 (4)	1965	2162	3800 (2)	4150 (1)
									Note 1		Note 1			
VIII *a* (tap)	038	.180	076	294	123	232	796	960			1900	2165	3150 (3)	3175 (2)
IX *e* (ten)	034	119	0[illegible]6	219	121	247	612	775			1800	2030	2925 (4)	2925 (1)
											Note 2		Note 2	
X *er* (pert)				331	131	239	570	712			1688	2188	3050 (2)	35[illegible] (1)
XI *a* (tape)	042	.172	091	305	125	235	494	614			3000	2800		
XII *i* (tip)	036	126	049	211	137	253	450	523			2950	2962		
XIII *e* (team)	036	189	116	341	136	252	296	332			2987	3266		1800 (1)
Means, or "Normals"	042	161	085	.288 (11) .296 (13)	125	244								

Note 1—Both of these sets of frequencies must be characteristic of *ar*. (Compare Fig. 13, also the results of Paget, quoted later.)

Note 2—The high frequency characteristics are less definitely located, for short *e*, than for any other doubly resonant vowel sound. (Compare Fig. 13.) The two sets of frequencies given above define a band of frequencies centered about 2,400 cycles within which the characteristic high frequency must be contained.

In the sounds of the first class the high frequency component is usually small in amplitude, more subject to individual bias in its frequency, and may or may not build up in amplitude as early as the low frequency characteristic. In the sounds of the second class the high frequency characteristic is usually prominent from the start and builds up very rapidly; while there is less variation in its frequency with the individual speaker. In sounds of the first class there is no decided suggestion of a transient in the high frequency (23 out of 40 records, Groups I to V inclusive) while in sounds of the second class the transient effect is pronounced (39 out of 40 records, Groups VIII, IX, XI, XII, XIII).

With these considerations in mind there is presented in table IV a summary of the data obtained from this preliminary examination of the vowel records. The mean duration time, and its subdivisions, are shown in the second column for each pure vowel sound, with mean duration only for the sounds *ar* (Group VII) and *er* (Group X). The fundamental and characteristic frequencies of each sound are shown in the 3 columns headed "Mean Fundamental," "Mean Low Characteristic" and "Mean High Characteristic Frequency" respectively. Each mean is taken from four records. The two columns headed "Scattered Low" and "Scattered High Frequencies" contain mean values of additional components, occurring in one or more records, in certain frequency ranges, the number of records in which such components are noted being shown in parentheses following the mean. The table illustrates and emphasizes many points which have been brought out in the preceding discussion, particularly the closeness with which the high frequency characteristics are defined in the vowels of the second or "doubly-resonant" class.

The table however gives no quantitative statement of the energy distribution among the different frequencies and it is necessary now to refer to the results of a harmonic analysis of these records which has been made and published[1] from which the diagram of Fig. 13 is taken. The machine method for analysing these wave-forms has been described by Mr. Sacia in detail elsewhere;[2] it suffices here to note merely the essentials in the treatment of the data.

For the dynamical study, the whole record from start to finish was taken as the unit for analysis, and the data obtained are therefore the average characteristics of the sounds throughout their duration. In the form of an endless belt each of these records was passed repeatedly through the analysing machine. A single record is of course

[1] "Dynamical Study of the Vowel Sounds." Bell System Technical Journal, III, No. 2, April, 1924.

[2] C. F. Sacia: "Photomechanical Wave Analyzer Applied to Inharmonic Analysis;" Jour. Opt. Soc. Am. and Rev. of Sci. Inst., 9, Oct., 1924, p. 487.

a non-periodic function, represented analytically by a Fourier Integral, not by a Fourier Series. The continued repetition of the record, 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, and it is this series of relative amplitudes at different frequencies which is given by the mechanical analysis of the records.

It would be possible to present these results as the sound spectra of the vowels, showing their original acoustic pressure amplitudes[3] but this treatment has been modified for practical reasons to take into account the relative importance of the various pitches in hearing. Using the available data on the relative sensitivity of the ear at different frequencies[4] the pressure amplitude at each frequency has been multiplied by the corresponding ear sensitivity factor and the resulting curves are taken as the *effective* amplitude frequency relations which are most generally characteristic of these sounds.

The data from the four male records and from the four female records of each sound are separately averaged and the resulting curves are shown in the diagram (Fig. 13). This averaging process was somehwat 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 on the musical pitch scale 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. Thus the final spectrum diagram 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,

[3] In Fig. 1, 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.

[4] See Bell System Tech. Journal, Vol. II, No. 4, October, 1923. The paper on Audition, by H. Fletcher, shows a graph of the "Threshold of Audibility" curve from which these data were obtained. The ear sensitivity factors used, of course, relate to the lower intensity levels; but it is thought that no essential inaccuracy is thereby introduced, as the position of the characteristic frequencies of a given vowel is subject to some variation with different speakers, and moderate variations in the height of these maxima in the energy spectra are not significant, except when taken from cycle to cycle in the case of an individual sound.

not the sizes of these curves which are significant. The order in which these curves are arranged is based upon the vowel triangle, and on Table IV. To return to the general discussion, we find that the fundamental voice frequencies do not have large effective amplitudes; it is interesting to note that these can be largely eliminated without impairing the distinctive quality of a vowel sound. The "scattered low frequencies" of the table (Sounds I to VII) exhibit appreciable amplitudes in the diagram. The "Scattered High Frequencies" of sounds I-VII previously noted exhibit small amplitude in the diagram. These are perhaps not essential to these speech sounds, but we should expect to find them in well trained singing voices. They are to a certain extent (particularly for the male voices), paralleled by the high-frequency regions of resonance for these sounds given in Paget's diagram, to which reference was made in Section I. Paget, it must be noted, is convinced that these high frequency regions of resonance are characteristic of the sounds of Groups I-VI.

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 most languages. The broad feature of the diagram is of course the progressive rise in frequency and gradual narrowing in range of the characteristic region of resonance, till the sound *a* is reached, succeeded by a splitting up into two regions of resonance which recede from one another as we follow the diagram downwards from *a* to the end. The exact location of sound X (er) is somewhat indeterminate, but it is evident that it belongs in the series of doubly resonant vowels. It is interesting to note that the distribution of the components of *ar* (refer either to Table IV or Fig. 13) is similar to the distributions given by Miller and by Paget for a form of the vowel *a* having "double" resonance; it is therefore as well located as any vowel in the series.

The characteristics of the *r* sound (whether considered as vowel or consonant) offer an interesting study, and in considering them we have an illustration of the practical value of records of the type shown. The problem of pronouncing a pure *r* sound is difficult; *r* is probably as variable in quality as any sound in the language, and it differs more than any other sound from one language to another. The precise location of its characteristic frequencies is thus a rather difficult matter. The records of *ar* and *er* disclose a noticeable tendency in speaking to make these sounds into diphthongs, the earlier portion of the record being nearly a pure *a* or (short) *e* while the latter portion of the record increasingly displays *r* characteristic. One speaker (MA) succeeded in making records for these two sounds which have nearly the same character throughout (Plates 49, 73), but for the other seven speakers, the "*r*" characteristics are best displayed toward the end of the record, though there is no sharp transition point. In the statistical study of these sounds the data were taken from the latter portions of the records; but in the mechanical analysis it was thought best to use the whole record. Now abstracting and condensing the data obtained in these two ways we have (ignoring fundamental tones) the following table of frequencies:

r (*ar* and *er*)

	From Table IV		From Fig. 13	
	Male	Female	Male	Female
Low	*570 630* 917 (ar)	*701 712* 1012 (ar)	*483 574* 861 (ar)	*512 542* 861 861
Middle			*1218 1448*	*1218 1448*
High.	*1088 1965*	*2162 2188*	*1933 2896*	1625 (er) *2435 2435*

These may be compared with Paget's results (from the second memoir, in which *r* is classified as a consonant sound) taking one of his general results from a mass of experimental data:

r (Paget: reference 9a, 9b p. 154)

"Throat or back resonance"	400-700 cycles
"Middle resonance"	1149-1824 cycles
"Front resonance"	1824-2169 cycles

(all varying with the associated vowel)

The *italicized* values in the first table above indicate correspondences with Paget's data, and we conclude that these roughly define the *r* sound, in terms of the steady-state theory.

Before taking leave of the vowel diagram, we should note 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 essential characteristics of the vowel sound. In other words, the individual vowel quality depends not only on a certain characteristic region of resonance but on the relative pitches in case there is more than one region of resonance. This effect is clearly shown to some degree in every group save one (VII:r) in Fig. 13. It will be noted that for the characteristic maxima of energy in the spectrum of a given sound, the peaks in the curve for female voices tend to occur at a higher frequency than the corresponding peaks in the curve for the male voices; but the musical interval between characteristic peaks for a given sound is about the same in the two cases. It is only in this way that we can account for 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.

V.

Four Semi-Vowel Sounds[1]

Now consider the sounds *l*, *ng*, *n*, *m*, which pronounced with the vowels *oo*, *ee*, *a*, following them, are arranged in Groups XIV and XV. Following the plan previously used, note first the general characteristics of these 24 records, made by the two male speakers MA and MB. An outstanding feature of the records is the diphthong quality which is clear in all: the transition is quickly made from semi-vowel to the affixed vowel sound and except in two records (Plates Nos. 108 (*lee*) and 113 (*ngee*) a definite transition point can be fixed. Marking this point for all records we find an average duration of 0.16 second for the semi-vowel sound, of 0.21 second for the vowel sound, mean total duration being 0.37 second. Noting the fundamental frequency in two locations, namely at the start and just before the transition point, it is found that there is a progressive rise in pitch during the record of the semi-vowel sound; this effect is in agreement with the individual characteristics of these two speakers previously noted in the pure vowel records. But in addition it is noted that the average fundamental for these two speakers (see Table V below) is somewhat below that previously used by them in the vowel records. (Refer also to Table III). This slight lowering of fundamental pitch may possibly be a characteristic of the semi-vowel sounds; and this effect occurs, as we shall see later, to a pronounced degree in the consonant sounds.

The amplitudes of these semi-vowel sounds are on the whole smaller than the amplitudes of the affixed pure vowel sounds, but some of them are surprisingly large. The low frequency characteristic of *l* is (for these voices) principally a third harmonic of the fundamental. With *n* and *ng* (which are nearly indistinguishable) the second harmonic becomes increasingly important, and in the *m* records it is very large. The high frequency characteristics of all four sounds lie between 2400 and 2900, falling somewhat as we pass through a sequence from

[1] A preliminary report has been made on the properties of these sounds, and their relation to the general vowel diagram. (Phys. Rev. 23, 1924, p. 309.)

TABLE V

Speakers' Characteristics, Semi Vowel Sounds

Sound	Duration in Seconds			Mean Fundamental (Semi-Vowel)	
	Semi-Vowel	Vowel	Total	At Start	Before Transition
l	16	20	36	100	107
ng	16	20	36	101	104
n	16	22	38	98	107
m	17	20	37	100	105
Mean	16	21	37	100	106

l to *m*. We have here, then, a group of doubly resonant sounds whose characteristic frequencies, whose amplitudes, and general behavior are such that they must be definitely related to the standard vowel diagram.

The amplitude frequency relations as obtained from a mechanical harmonic analysis, and corrected for the variation in sensitivity of the ear are shown in Fig. 14. The process of mechanical harmonic analysis has been outlined in connection with the vowel records, and the procedure was the same here, except that only the semi-vowel portion of the records was taken as the unit for analysis. The record for analysis was cut at the end of the last cycle before the transition point, and two profile copies of the semi-vowel wave were joined together in an endless belt which was passed through the analyzing machine.

Aside from the close resemblance between the frequency spectra of the four sounds the noteworthy feature of Fig. 14 is in the similarity between the *l* spectrum and that for *ee* as previously given in line XIII of Fig. 13. The essential differences are a slight increase in the importance of the low frequency characteristics, and the slight shift of all the resonant regions toward lower frequency, in passing from *e* to *l*, and on through the sequence *ng*, *n*, *m*. We may thus regard the chart of Fig. 14 as a logical continuation of the generally accepted chart of Fig. 13 and place the four semi-vowel sounds definitely in an extended vowel diagram, following in regular order the sound long *e*.

Sir Richard Paget has made the interesting statement that "all the consonant sounds are as essentially musical as the vowels, i. e., they depend on variations of resonance in the vocal cavity, and should be capable of being imitated in the same way, if their characteristic resonances could be identified and reproduced in models." It is interesting to compare some observations made by him on *l*, *ng*, *n*, *m*, and reported in his second memoir. Working according to the method previously described (§1) Paget has constructed resonators which, under certain conditions, will produce transient forms of the four sounds we are discussing. Their tone constituents are identified by him as follows:

RESONANT FREQUENCIES, SEMI-VOWEL SOUNDS

(Paget: Reference 9b)

	"Throat"	"Middle" (Nasal)		"Upper" (Oral)
l	228-406[1]	683 (faint)		1625-1932[1]
n	203-228	683	1217-1366	1448-2169[2]
ng	203-228	541-724	1217-1448	2298-2579
m	271	. .	1217-1448[2]	861-1722[2] 2434-2579 (faint)

[1] Varying and finally approximating a characteristic region of resonance of the associated vowel.

[2] Varying with the associated vowel.

Studying Paget's results in connection with those of Fig. 14, we note that the energy spectra clearly show the "throat" resonances for all four sounds in the neighborhood of 256 cycles. In the case of *n* the nasal resonance at 683 cycles (Paget) is one of the prominent tones centering around a frequency of 512 in the spectrum diagram. This resonance also appears prominently in the spectrum for *m* though Paget did not notice it. The higher middle resonances (1217-1448 cycles) which appear in Paget's table for the last three sounds appear also in the spectra for these three sounds according to Fig. 14. Allowing for the variation stated in notes (1) and (2) above, it appears that the upper (oral) resonances for the four sounds, as noted by Paget, are essentially the same as those that appear in all four spectra in the diagram in the range of 2048-2896 cycles.

With regard to Paget's observations on the transient character of these sounds (he classifies them as consonants) and on the variability of some of their components (Notes 1 and 2 of table above), depending on the associated vowel, there is room for some difference of opinion and the reader may form his own conclusions after a detailed inspection of the records shown. Taking the sound l for example, and studying first the three records *loo*, *lee*, *la* by M A and then the three corresponding records by M B it seems to the writer that such variations as are noted in characteristics are due not so much to change in the associated vowel as to the change in the speaker, and a similar conclusion will probably be reached for each of the other three semi-vowel sounds.

From the evidence in the records, it is difficult to subscribe entirely to a "transient" theory of these sounds, at least when they precede the standard vowel sounds. The evidence justifies the use which has been made of the steady-state idea, and the harmonic analyses leading to a determination of characteristic frequencies. But there is a possibility that the harmonic analysis does not tell the whole story. These two groups of records and the acoustic spectra based on them furnish outstanding examples of the niceties involved in speech and hearing in order to achieve the miracle of articulate speech. Without harmonic analysis, the most casual observer will note, for example, the similarity between the corresponding records of the *l* and *n* sounds, but more astonishing still is the resemblance between the *l* and *ee* sounds shown together in Plates Nos. 107 and 108. In this latter case (*l* and *ee*) practically the same high and low characteristic frequencies are involved, and it would seem that the distinction, which is sufficiently pronounced to the ear, must be based to some extent not only on the relative amplitudes of these frequencies present, but also on the behavior of these amplitudes during the fundamental cycle. It will be noted in practically all of the records of these semi-vowel sounds that the high frequency characteristic is a transient of more rapid decay than in the case of the pure vowel sounds; it is not of large amplitude except at the beginning of the cycle. On the face of the records this is the only explanation available for whatever distinctive quality these sounds, as a class, must possess.

VI

Sixteen Consonant Sounds

The last two groups, XVI and XVII contain, respectively, records of the "hard" and "soft" consonant sounds, each with the *a* sound affixed, and pronounced by the two male speakers. Here the classification is somewhat arbitrary; it is difficult if not impossible to arrange the sounds of these two groups in any such satisfactory series as has been determined for the semi-vowels of the two preceding groups. The sounds *dth* (that) and *th* (thin) for example have transitional characteristics that relate them to both groups; but they are placed at the end of Group XVI, to emphasize their relation to the pair *v f* of the last group. With these reservations as to arrangement, consider the general characteristics of the consonant sounds of these two groups.

Examination first discloses a relatively easy separation of a given record into a consonant and a vowel portion and, as might be expected, a longer duration for the "voiced" consonants. In all the voiced consonants a sufficient portion of the record is reproduced to illustrate the voicing or fundamental of small amplitude in the early stages of the record; in the case of the unvoiced consonants of Group XVI this is not necessary. In the case of both the voiced and unvoiced consonants of Group XVII, longer records are shown, the high frequency component making this necessary, although the fundamental does not appear in the early stages of the unvoiced consonants of this group. The mean duration of the voiced consonants (*b*, *d*, *g*, *dth*) of Group XVI is 0.14 second; of the unvoiced consonants (*p*, *t*, *k*, *th*) 0.05 second. Aside from traces of the fundamental tone (and traces of its second and third harmonics) there is nothing of interest in the early stages of three of these four voiced consonants; in the case of *dth* there are traces of a high frequency (4200 and 2600 in the two records) in the early parts of the fundamental cycle. The voicing for all four sounds if uniformly of lower pitch than that used later in the records in speaking the vowel sound. Leaving the early stages, the record then proceeds to a transition point, lasting through from one to four cycles of the fundamental, and culminating in the appearance of the vowel sound. Before this transition point is reached, traces of high frequency appear in most cases, sometimes suggesting a single transient vibration. Aside from the lack of the fundamental vibration, there is a further distinguishing characteristic of the "unvoiced" sounds: a tendency of the first transition cycle of the fundamental to appear from 10 to 20 per cent shorter in duration than the mean of several following cycles. With both voiced and unvoiced sounds there is a tendency for a moderately low frequency (500 to 700 cycles) to appear during the transition; also a high frequency (of mean value 3225 cycles for the 16 records of this group) which latter may be due to the beginning of the *a* sound. Some of the individual characteristics of these records are given in Table VI.

The notable distinction between these sounds and the sounds of the next Group (XVII) rests on duration factors, and of even more importance, the pronounced high-frequency characteristics of the sounds of the last group. The mean duration of the voiced sounds in Group XVII is 0.21 second; that of the unvoiced sounds, 0.18 second. Two of the other characteristics are similar to those noted in the preceding group; first the voicing, where it occurs, is of abnormally low frequency, and second in the case of the unvoiced sounds, there is a marked shortness of the first fundamental cycle at the transition point. Except in the case of the sound *v* (Plates 145 and 146) the high frequencies are persistent and in many cases of large amplitude, both at the start and during the course of the consonant sound. These frequencies rise, as we go through this group, to values of 7000 and 8000 cycles in the case of the sounds *z* and *s*, shown in the last four records. For a full appreciation of these pronounced high frequency characteristics reference must be made to the records themselves, or the summary of characteristics, in Table VII. Here again, in distinguishing these sounds the remarkable performance of the ear is manifest, and the recording apparatus is used nearly to the limit of its utility.

We may best conclude this discussion of the consonant records by brief comments on some of the individual sounds, and a comparison where possible with data given for them in Paget's second memoir.

B P.—(Plates 129-132). Both Paget (ref. 9b, p. 165) and Miller (ref. 3) have noted the essential impulsive quality of these sounds, and have produced them by sudden closing and opening of the mouth of a resonator. Paget considers *p* to be the more suddenly released, i. e. to have the steeper wave-front. From the records this is not evident; following the voicing period, the *b* would seem to be more suddenly produced, as judged by the growth in amplitude of the *a* sound following.

D T.—(Plates 133-136). For both of these (see either Table VI or the records themselves) we note a high frequency characteristic of about 4000 cycles. Paget (9b, p. 168) observed "an upper resonance 5 to 8 semitones higher than that of the associated vowel, and a low resonance of about 362." We note in the records a low frequency of the order of 500 in the case of *d*. Paget notes a "greater amplitude in *t* due to higher air pressure" and the records show a greater amplitude for the high frequency in the case of *t*, except right at the transition point, where *d* shows the high frequency of large amplitude. No conclusion can be given as to relative steepness of wave-front, *d* vs. *t*, because in both cases we note for speaker MB (Records 134, 136) a steeper wave-front than for MA (Records 133, 135). The difference between *d* and *t* may depend entirely on the voicing and on the complicated phenomena at the transition point.

G K.—(Plates 137-140). *k* shows the characteristic transients (1500, 4000; Table IV, notes 4 and 5) to much more pronounced degree than *g*. From the records it would seem that *g*, in addition to the voicing, disclosed a steeper wave-front, the *four* transitional cycles required for *k* (records 139-140) emphasizing this point. No other generalizations seem warranted, on account of the complicated series of events recorded. These sounds are treated at length by Paget (9b, p. 171-173) who observes considerable variation in their resonant ranges, depending on the associated vowel. It will be noted however, that in these four records particularly, consonant characteristics are persistent and of large amplitude before the vowel sound begins to appear.

DTH/TH.—(Plates 141-144). The high frequencies (2600, 3000, 3200) culminating at the transition point seem to be the key to these records. They are more persistent for *dth*, while *th* appears to show the steeper wave-front. Paget states (9b, p. 158) that "in δ [*dth*] the middle resonance [1149-1932, his figures] is overblown, - - - louder than the corresponding resonance in θ [*th*]." He gives also an "upper sibilant of 3444-5950," louder for *dth* than *th*, and "difficult to identify." It will be noted that in one record for *dth* (no. 141) there is during the voicing period a faint high frequency which has been set down in Table VI as 4000 cycles. This faint "sibilant" (which may always be audible though it fail to be recorded) establishes a certain kinship between these two sounds and those following (the fricative consonants) which are rich in sibiliant sounds.

V/F.—(Plates 145-148). *v* shows a pronounced voicing, and as previously noted, a less prominent high frequency component than its partner *f*, or any of the other fricative consonants. Comparing *v/f* with *dth/th* it seems from the records that the former pair are of higher frequency (particularly *f*) and that for *v/f* as a unit the high frequency characteristic is more pronounced; just the opposite conclusion to that reached by Paget (9b, p. 161-162). *f* may indeed differ more from *v* than *v* from *dth*, thus raising difficulties of classification both physically and phonetically, which cannot be resolved on the basis of the few records available. The exceedingly fine distinction between the sounds *v* and *dth* could be no more strikingly shown than it is in the records given, for both speakers.

J/CH.—(Plates 149-152). Some of the recorded phenomena of this pair suggest correspondences between them and the pair *g/k*; but the pair *j/ch* shows a higher frequency characteristic during the important mid-portion of its history. Of the pair, *ch* seems to show the steeper wave-front, that is, the more rapid transition to the vowel sound.

ZH/SH.—(Plates 153-156). With this pair we pass to the field of pure sibilants, in which there is no evidence of impulsive action or steepness of wave-front. The action seems to be that in the voiced sound, there is, in addition to the presence of the fundamental tone, a breaking up of the characteristic high frequency wave-train into discrete units corresponding to the fundamental tone, whereas in the unvoiced sound the high frequency characteristic is continuous, though irregular. Thus noting that the characteristic frequency is of 3000 to 1600 cycles the outstanding phenomena of *zh sh* are well defined. In addition to frequencies of 2048-3219 noted by Paget (9b, p. 163) he gives a "pronounced middle resonance of 1625-2048." This latter observation of Paget's may correspond to the 1800-2000 frequency in the records of MB (Plates 154, 156) in the transition region, but this component does not seem to be prominent in the records.

Z/S.—(Plates 157-160). The general properties of these sounds can be inferred from the discussion of the preceding pair (*zh/sh*), adding only the fact that their principal characteristic is of much higher frequency. From Table VII we note a range of 1200-8000 cycles; Paget (9b, p. 162) gives "a characteristic upper resonance of 5790-6886." Paget also gives "a middle resonance of 1084-2298." The records do not show as low a range of characteristic frequencies unless it be the frequency range 2200-2800 (see Note 1, Table VII), within which fall certain vibrations occurring in the early parts of the fundamental cycles of the voiced sounds *zh* and *z*. The true *s* sound is, as Paget has stated, "a relatively complex hiss" and this is true of *sh* as well. And to complete the record, we must observe that *zh* and *z* are even more complex, if possible, and thus not inappropriate examples of the sounds of speech with which to conclude this survey.

To summarize, we have considered some of the more outstanding features of the wave forms of speech sounds which have been recorded. Many more detailed properties of these records deserve further study. The progressive change in wave form from cycle to cycle of the fundamental, particularly at the beginning of a sound, is undoubtedly an important factor in determining the character of speech sounds; it becomes most important, as we have seen, in the study of the more impulsive consonant sounds. There is material in these records for extended studies of this kind, which require a harmonic analyzer of a large number of components. We have not dealt with the question of the inherent power in speech sounds, another very characteristic property; these important data are accurately given in a paper by C. F. Sacia in this issue of the Journal. The relative power in consonant and vowel sounds can also be determined from those records in which vowels and consonants appear in combination, and it is hoped to carry this study further. Many other investigations of speech are now made possible on the basis of the accuracy of this set of records; in conclusion we may emphasize the fact that, for the present, the record is the important thing, and we believe that a set of faithful records opens a new prospect in the field of speech investigation.

Plate No.	Sound	Speaker	Duration	Consonant Characteristics: Near Start: Voicing (Fundamental and Harmonics)	Near Start: High Frequency	Mid Portion to End: Voicing	Mid Portion to End: High Frequency	Transitional Characteristics: Low Frequency	High Frequency (Note 6)	No. of Cycles	First Cycle Short	Vowel Fundamental: Near Start	Near End
129	*ba*	MA	12	90,180	none	90,180	none	700	2700	1		100	115
130	*ba*	MB	19	100,200	none	92,181	none	700	3100	1	yes	116	107
131	*pa*	MA	02	unvoiced	none	unvoiced	2800 (Note 2)	1000	3600	1	yes	100	111
132	*pa*	MB	04	(one 60 cycle vibration)	none	(One 60 cycle vibration)	3800	900	3600	1	yes	119	114
133	*da*	MA	13	90,180	none	79,158	3800 (Note 3)	500	2800	3	yes	103	115
134	*da*	MB	10	98,196	none	98,196	3600	600	3200	2		112	109
135	*ta*	MA	07	unvoiced	none	(One 100 cycle vibration)	4300 (Note 3)	. . .	3200	4	yes	101	112
136	*ta*	MB	06	unvoiced	none	unvoiced	3600	900	3000	2	yes	120	113
137	*ga*	MA	.12	100,200,300	none	81,252	1600, 2800 (Note 4)	550	3000	3		101	111
138	*ga*	MB	.10	100,200,300	none	95,190	1400, 1000	600	3600	2		112	112
139	*ka*	MA	.07	unvoiced	none	unvoiced	1500, 4000 (Note 5	1200	3800	1	yes	109	118
140	*ka*	MB	.08	unvoiced	none	unvoiced	1600, 4200	1300	4000	1	yes	125	116
141	*dtha*	MA	20	83,166	4000 (Note 1)	95,189	1200 (Note 1)	600	3000	2	.	101	116
142	*dtha*	MB	18	100,200	2600	100,200	2700	600	2600	4		109	107
143	*tha*	MA	.02	unvoiced	none	unvoiced	none	600	3200	1	yes	110	110
144	*tha*	MB	02	unvoiced	none	unvoiced (one 100 cycle vibration)	none	600	3200	1	yes	113	107

NOTE 1—A trace of these at beginning of the early fundamental cycles.
NOTE 2—One faint transient. NOTE 3—Transients; longer for *ta* than for *da*.
NOTE 4—One transient. NOTE 5—Irregular transients.
NOTE 6—Possibly due in some cases to the *a* sound.

TABLE VII

Group XVII — Fricative Consonants

Plate No.	Sound	Speaker	Consonant Characteristics					Transitional Characteristics				Vowel Fundamental	
			Duration	Near Start		Mid Portion to End							
				Voicing Fundamental and Harmonics)	High Frequency	Voicing	High Frequency	Low Frequency	High Frequency (Note 3)	No. of Cycles	First Cycle Short	Near Start	Near End
145	*ra*	MA	.20	97,195,390	3000	87,174	none	600	2700	3	...	101	116
146	*ra*	MB	.25	112,221	3200 (trace)	100,200	none	600	3100	2	...	112	107
147	*fa*	MA	.15	unvoiced	3100	unvoiced	3500, 7000	500	2800	4	yes	112	121
148	*fa*	MB	.30	irregular	3200, 6100	unvoiced	3200, 6400	600	3600	3	yes	111	104
149	*ja*	MA	.22	81,213	3400	81,162	2600, 5200	450	2700	4	...	110	110
150	*ja*	MB	.14	trace	3300	90,179	2000, 1800	500	3100	4	...	115	111
151	*cha*	MA	.07	unvoiced	4800	unvoiced	2800, 1800	{500, 1500}	3000	2	yes	101	111
152	*cha*	MB	.08	unvoiced	3600	unvoiced	3600, 6400	{500, 1600}	trace	2	yes	119	115
153	*zha*	MA	.28	86,172,341	3000, 1000 (Note 1)	87	3000, 1000 (Note 1)	450	2900	4	...	100	111
154	*zha*	MB	.13	96	2600, 1200	99	3000, 1200	{500, 2000}	....	1	...	111	111
155	*sha*	MA	.18	unvoiced	2800, 3600 (Note 2)	unvoiced	2800, 1600 (Note 2)	450	3200	3	yes	104	101
156	*sha*	MB	.17	unvoiced	2200, 5000	unvoiced	2600, 500	{500, 1800}	2800	3	yes	117	112
157	*za*	MA	.24	96,384	2800, 5600 (Note 1)	89,178	5200, 7000 (Note 1)	100	3100	4	...	98	108
158	*za*	MB	.22	100,300	2200, 4400	100,200	2800, 5600	550	2800	5	...	111	107
159	*sa*	MA	.27	unvoiced	5600, 8000	unvoiced	6000, 7800	500	2900	2	yes	114	111
160	*sa*	MB	.19	unvoiced	4000, 6100	unvoiced	4200, 6600	650	2900	2	yes	117	108

Speech Power and Energy

By C. F. SACIA

Introduction

In the past, much research has been devoted to the determination of the relative magnitudes of the frequency components of speech, and the results of these explorations are useful and well known. Thus the communication engineer is apprised of the frequency range over which his apparatus should respond uniformly in order that the transmitted speech suffer no frequency distortion. But to provide against load distortion, he requires the knowledge of a different kind of data: numerical values of the magnitude of power involved in speech waves as a whole. This investigation deals with the magnitudes and forms of speech waves primarily in terms of power, and is not concerned with frequency as the argument.

Although the subject matter is not fundamentally new, this treatment of it is somewhat of a venture. The broad classification of power is a convenience here, but its future value will be dependent upon engineering usage. I have also introduced the use of the peak factor, which, being a simple index of the wave form, may perhaps find application in vowel study and phonetics as well as in the technical field. A condensed table of peak factors was incorporated in Mr. Fletcher's compilation in the preceding issue of this Journal.

Derivation

The nature of power in a syllable of speech may be most easily comprehended by reference to an illustration such as that shown in Fig. 1. The representation of the instantaneous power (P_i) is an enlarged copy of a power oscillogram of the word "quite." Because of its extreme jaggedness, the curve had to be represented by a profile rather than by an outline. Although this is a quickly spoken syllable it plainly displays a cyclic repetition; the cyclic interval (for example, from *a* to *b* in the figure) is ordinarily called the vocal period and its reciprocal, the vocal frequency[1]).

One feature of interest may be noted here: the irregularity in the growth and decay of the peaks. This is evidence of a slight vocal

[1] The power due to any periodic force, containing only odd harmonics, fluctuates with double the frequency of the fundamental; but in the case of any periodic force containing even harmonics also, the power fluctuations have the same fundamental frequency as the force. Although speech sounds are not periodic an analogous relation exists for them.

tremolo. Tremolos usually occur in singing voices and vary widely in their character. They constitute modulations which in actual singing sometimes occur as slowly as two per second. The slower modulations affect the ear as beats or pulses, while the most rapid ones affect the quality by the resulting sidebands of overtones. Those

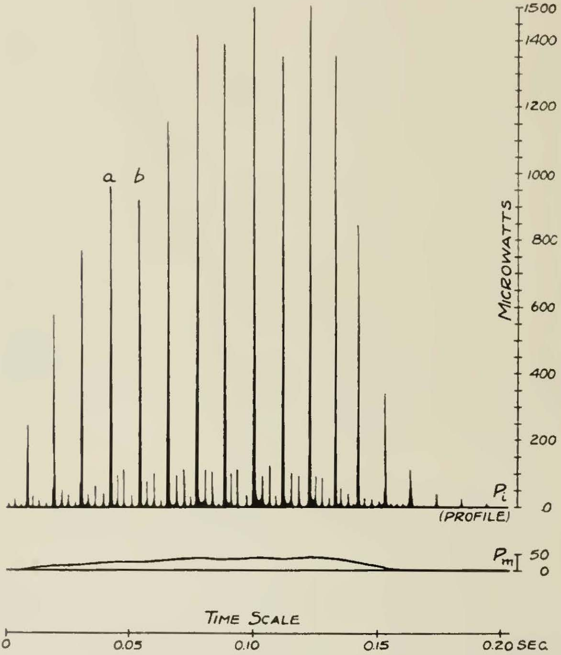


Fig. 1—Instantaneous and mean power. Enlarged copy of original oscillogram of the word "QUITE"

shown in the figure are of the latter types, their modulating frequency being about 50 per second.

From the instantaneous power we derive the mean power, P_m, whose chief significance lies in the fact that it is the kind of power that would be read by a quickly acting wattmeter; it is likewise proportional to the deflection shown by the ordinary a.c. voltmeter or ammeter, or by the volume indicator. A graph of the mean power may be obtained by drawing the average power in each vocal cycle and then drawing a smooth curve through the resulting broken line. This would be an impracticable way of obtaining curves of mean power; actually they have been obtained independently of the P_i curves in this work, in a manner described later.

Vowel sounds carry by far the most of the power and energy of speech, and it was to them that the above considerations were tacitly applied; but the definition of the mean power is similarly applicable to the semi-vowels, voiced consonants, and fricative consonants.

The peak factor is the square root of the ratio of a peak value of P_i to the corresponding value of P_m.

Still another commonly used interpretation of power is made in terms of its average over an entire syllable, word or speech. Such an average, although the same for instantaneous and mean power, is most easily determined by means of the latter: it is the total energy divided by the time involved. Graphically it is the area of the P_i or P_m curve divided by the base. If the base includes the silent intervals between syllables the result will be called the long average; if the silent intervals are excluded from the base, the result will be called the short average.

Thus it is seen that the word "power" when applied to speech has a variety of meanings and always needs to be qualified. For example, the speech of a certain person may have shown a long average power of 10 microwatts while the instantaneous power frequently rose to 2,000 microwatts.

In obtaining the power, we obtain indirectly the pressure on the condenser transmitter, which is located 9 cm. from the speaker's lips. In the treatises on acoustics, the power of a simple-harmonic wave is derived in terms of the pressure,[2] the numerical result being at 20° C,

$$P = \frac{p^2}{415} \tag{1}$$

where P is the power in microwatts across 1 sq. cm. of wave front, and where either mean or peak value is taken for both power and pressure. Here we are not concerned with simple harmonic waves, but the same result holds for instantaneous, mean, or average values in any kind of wave, since

$$P = \frac{1}{10} p \frac{d\xi}{dt} \text{ microwatts across 1 sq. cm.,}$$

[2] See, for example, Rayleigh: Theory of Sound, Vol. 2, page 16.

and the air particle displacement,

$$\xi = \frac{1}{41.5}\int p\,dt \qquad \text{(41.5 is a resistance factor)}$$

for a wave travelling in the positive direction.

From the power intensity thus found at the transmitter we can obtain an estimate of the power developed by the speaker. With the transmitter surrounded by a plane reflecting surface so as to give reflection for speech frequencies, the pressure is doubled and the power intensity quadrupled over the values they would have in free air, hence the observed intensity is divided by 4. The usual assumption is made that this same intensity is distributed over a hemisphere whose center is at the speaker's lips. Hence the required estimate of the speaker's power is obtained by multiplying the measured power intensity at the transmitter by the factor $\frac{\pi 9^2}{2} = 127$. For the sake of convenience, these two values are always given together in the accompanying tabulated results.

Instantaneous and Mean Power

In dealing with the power in a syllable, the matter of greatest interest is the maximum values attained by P_i and P_m throughout the entire syllable. These maxima will be denoted by $\overline{P}_i$ and $\overline{P}_m$, respectively. Table I shows their approximate ranges in the case of accented syllables.

TABLE I

Instantaneous and Mean Power

Typical Maximum Values for an Accented Syllable

	Speaker's Power Microwatts	Power Per Cm.² at Transmitter
$\overline{P}_i$	1000 to 2000	8 to 16
$\overline{P}_m$	60 to 120	0.5 to 1.0

At this point it is worth while to consider an application of the foregoing. A salient characteristic of speech waves is the generally high ratio of peak value to mean square value (peak factor), as can be inferred from Fig. 1. Failure to take this into account frequently causes load distortion in speech transmitting amplifiers. It sometimes happens that the effective output voltage or current has been measured, and the assumption of an equivalent sine wave (i.e., one having the same effective value) is made; but this leads to a large error in the estimate of the peak value. Thus with an insufficient allowance made for the peak voltage impressed upon the grid of the tube, there is the possibility of the grid becoming momentarily positive due to insufficient negative bias or still worse, the plate may be overloaded by the peaks. The resulting suppression of the peaks in the sound output can readily be detected by an accustomed ear, provided that the whole system is reasonably free from frequency distortion.

Average Power

In Tables II and III are summarized the observations made upon the two speeches which were used in this work. There are two reasons for showing them separately: the two speeches were not spoken in immediate succession; and they differ somewhat in character, the first being declamatory while the second is of a more conversational nature. This difference is not very great, but should account nevertheless, for the slightly higher values in Table II. By taking the weighted mean of the first number in both tables, we obtain 7.4 microwatts as the long average power in normal speech.[3]

TABLE II
First Speech, 50 Syllables
Average Power in Microwatts

	Long Average		Short Average	
	Speaker's Power	Per cm^2 at Trans.	Speaker's Power	Per cm^2 at Trans.
Composite of 16	8.6	0.067	13.1	0.102
Composite of 8 male	8.2	0.064	12.7	0.099
Composite of 8 female	9.0	0.070	13.5	0.105
Maximum male	10.6	0.082	17.1	0.133
Maximum female	17.0	0.131	21.8	0.169
Minimum male	7.0	0.055	10.8	0.084
Minimum female	5.7	0.044	8.8	0.069

TABLE III
Second Speech, 72 Syllables
Average Power in Microwatts

	Long Average		Short Average	
	Speaker's Power	Per cm^2 at Trans.	Speaker's Power	Per cm^2 at Trans.
Composite of 16	6.6	0.054	9.9	0.080
Composite of 8 male	6.2	0.050	8.9	0.072
Composite of 8 female	7.1	0.057	10.8	0.087
Maximum male	8.1	0.065	13.0	0.105
Maximum female	9.8	0.079	15.1	0.122
Minimum male	3.9	0.032	5.7	0.046
Minimum female	4.0	0.033	6.0	0.048

Note: The average ratio of the total time in the silent gaps to that consumed by the syllables is 0.55; the syllables average 0.16 sec.

[3] Crandall and MacKenzie gave an estimate of 12.5; B. S. T. J., Vol. 1, No. 1; Phys. Rev., Mar. 1922.

Stress

Since our observations have shown qualitatively that the louder syllables have the greater rise of mean power, means are available for calibrating the stress modulation of the voices under test. To form a discriminant for each speaker we proceed in the following way:

(1) Measure the $\overline{P}_m$ of each syllable;
(2) Find the ratio of each $\overline{P}_m$ to the greatest $\overline{P}_m$ occurring in the speech; call this ratio ϵ;
(3) Find the proportional number, $s/\underline{s}$, of syllables for which ϵ is greater than the magnitude n, where n may vary between 0 and 1;
(4) Plot the variables $s/\underline{s}$ and n against each other to give the required curve.

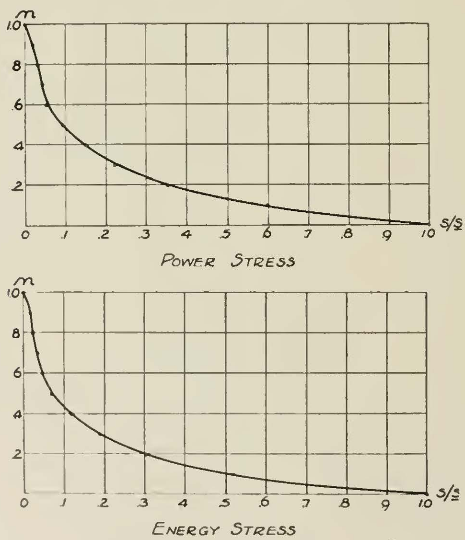


Fig. 2a—Composite stress curves of 16 voices

The analogous relation between syllabic energy and stress is found by using the total energy of each syllable instead of $\overline{P}_m$ in the above.

A large number of these curves has been so obtained, but it will suffice to consider here a few of the representative types. Fig. 2a shows composite curves and Fig. 2b gives a series of each kind of curves for four speakers. Note the changing mode of stress which is shown in the sequence from top to bottom: in the first case the syllables of weaker stress greatly predominate while in the last case there is a more nearly uniform distribution of the syllables with respect to the degree of stress. It is evident from a comparison of the two series

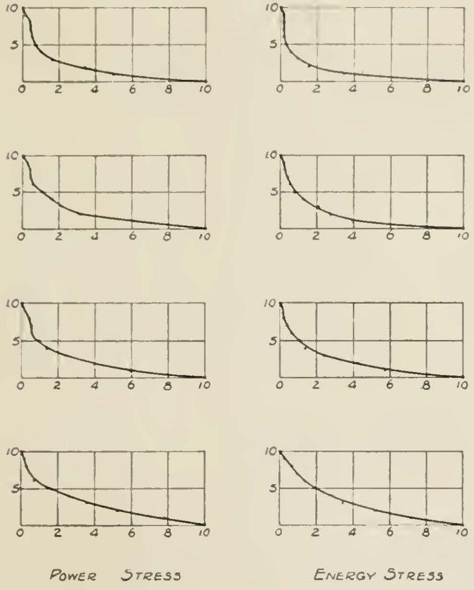


Fig. 2b—Types of stress curves

that the speaker's type is much the same whether judged by the power or energy standard. An exceptional case might arise, however, if one should put emphasis on a syllable by prolonging the time of utterance, for here the increased energy of the syllable would not necessarily mean a greater stress. But from the point of view of phonetics, the energy method should be useful in calibrating emphasis, which can be taken as a function of time of duration as well as of mean power.

Relative Power of Vowels

One test which was made on the speakers was for them to utter disconnectedly and without accent eleven monosyllables, each of which contained a fundamental vowel sound. The results of this test give a general indication of the inherent power, $\overline{P}_m$, in unaccented (but unslighted) vowels relative to each other. The difference between the

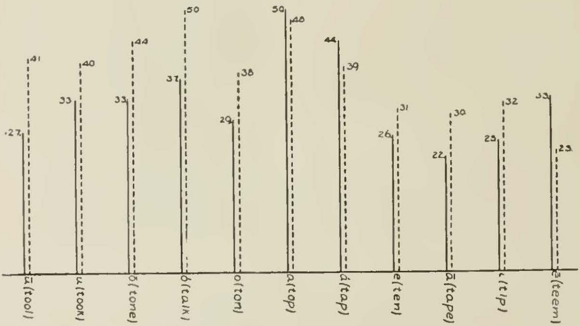


Fig. 3—Inherent relative power

——— Indicates Male Voices ------ Indicates Female Voices

Numbers indicate approximate power from voice (in microwatts)

male and female voices in this respect warrants separate charting of these characteristics. Fig. 3 shows the chart in which the vowels are arranged in the sequence[4]) the first half of which accompanies an increase in the angle of the speaker's jaws, and the succeeding half accompanies an increase in the elevation of the tongue.

It might have been anticipated that the more open vowels have more power; but there is apparantly one irregularity in this tendency in the case of the vowel o (as in ton). Furthermore, the vowel ē (as in teem) looks somewhat different for the two voices, when compared with the vowels immediately preceding it in the series. There is some difficulty in uttering it so as to make it carry, in the case of female voices a fact which I have previously encountered when recording them. The male voice, on the other hand, shows a decided rise in this direction. The advantage in the case of ū (tool) is reversed: here the male voice begins to fall off while the female voice stays about the same. These results suggest a difference in the resonant structure

[4] This arrangement is based upon the well known vowel triangle of Victor.

between the male and female voices, which, however, does not affect the higher frequencies enough to alter the vowel characteristics.

Peak Factor

The tests just described were also used to obtain the peak factors of the vowels. These were determined by measurement of the maximum P_i and P_m of each syllable and are charted in Fig. 4. Here again there

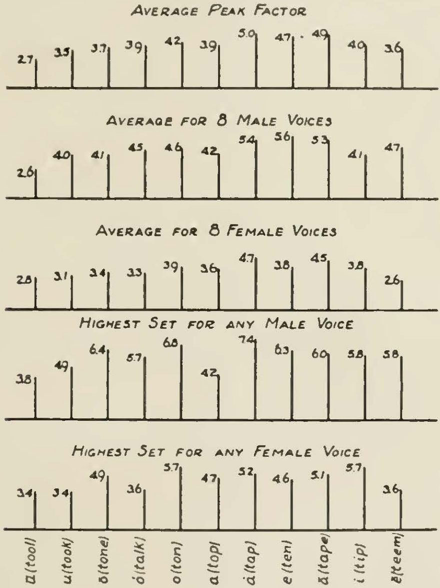


Fig. 4—Peak factors of vowels

are differences between the sets for the male and female voices, the former being somewhat higher, especially for the vowel ē. In both cases such rasping vowels as á (tap), e (ten), ā (tape) have sharp waves and high peak factors. Having listened attentively to all these voices under test, I have become able to associate peak factors with vocal qualities in the following way: the voices with the higher peak factors are those which in the ordinary terminology are said to be "resonant" or "vibrant"; they have the greater carrying power, especially over the telephone; they are rich in the musical sense and are therefore well suited to singing, although many such voices, unfortunately, are never applied to the art.

To illustrate an application of the peak factor to engineering, we shall again take into consideration the speech amplifier whose mean effective output voltage is indicated by a suitable device such as a volume indicator. From this, the peak value of the instantaneous voltage is wanted; to find it necessitates a knowledge of the peak factor. Now since the latter differs somewhat for different sounds and speakers, it is necessary to use one factor which makes allowance for the worst cases (highest voltage peaks) which can occur often. For most purposes, the factor 5 will suffice, hence the rule is: the mean effective voltage should not exceed one-fifth the overload voltage of the system.

Apparatus

In order that the apparatus (see Fig. 5) be a faithful recorder, it was made with the following characteristics:

(1) A nearly distortionless reproduction of wave form by the condenser transmitter and amplifier.

(2) A full-wave parabolic rectification of the amplifier output.

(3) Load capacity sufficient to transmit the high sharp peaks of speech waves without cutoff.

(4) Uniform response, from 0 to 6000 cycles in the oscillograph vibrator recording instantaneous power.

The calibration of the amplifier and condenser transmitter is shown in Fig. 6. To make the overall characteristics so nearly uniform it was found necessary to use the resonant circuit in the output of the second N tube, this compensating for an irregularity due mostly to the 45 feet of cable which leads from the transmitter and first stage of amplification in the sound-proof room to the main part of the amplifier.

The oscillograph (see Fig. 5) was provided with two series connected vibrators one of which was sensitive to low frequencies only, and recorded the mean power. Although it did not completely suppress the fluctuations of vocal frequency, it reduced them to the order of small superimposed ripples through which the P_m curve could be drawn. The instantaneous power was recorded by the other vibrator whose characteristics are noted in item (4) above.

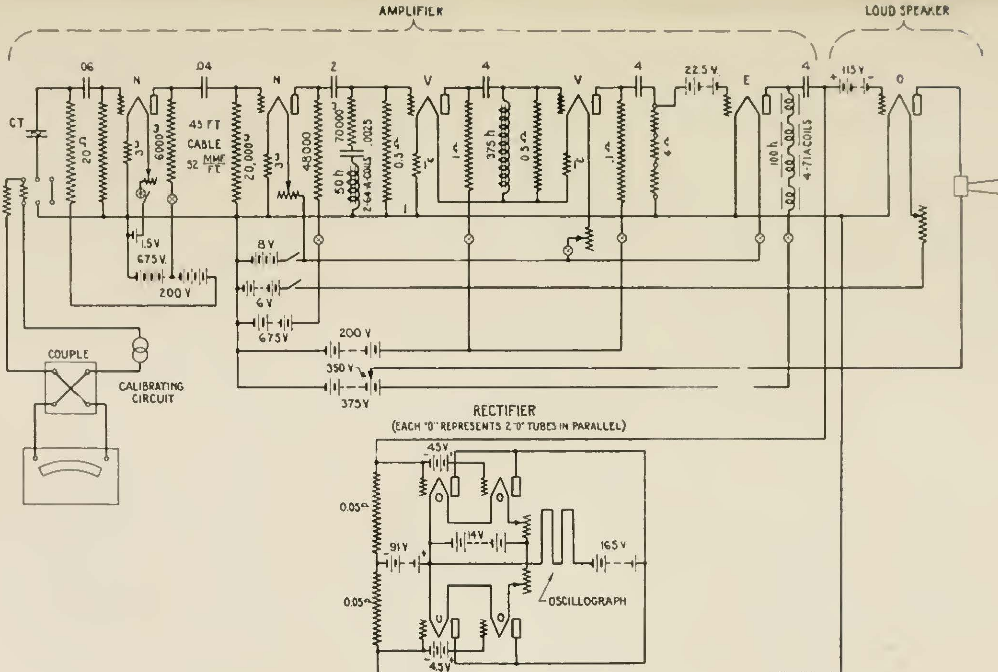


Fig. 5—Speech power recording circuit

TABLE IV

Calibration Constants

(a) Constants of Vibrators $I/D =$
 (1) Low frequency .... 5 milliamperes per cm.
 (2) Instantaneous power .... 286 milliamperes per cm.
(b) Rectifier constant $E^2/I = /40$ (volts)², milliamp.
(c) Pressure on transmitter vs. amplifier output $p^2/E^2 = 1/2.95^2$ dynes²/cm⁴ volt².
(d) Power intensity at transmitter vs. pressure $P/p^2 = 1/415$ cm² microwatts/dynes².

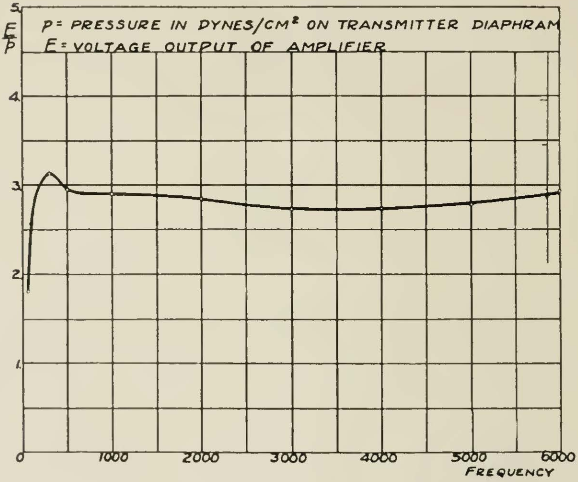


Fig. 6—Calibration of condenser transmitter with amplifier

The product a b c d gives $P_m/D_m = 0.192$ microwatts per sq. cm. of wave front as indicated by a deflection of 1 cm. of the oscillograph low frequency vibrator. Similarly $P_i/D_i = 11.1$ for the instantaneous power vibrator.

Method

Records were made on sensitized paper strips 6 cm. wide moving at a velocity of about 20 cm. per second. Three graphs were traced simultaneously, the instantaneous power, the mean power, and the timing wave of 100 cycles from an oscillator. When connected speech was being recorded, the oscillograph operator listened to the speech as reproduced by the loud speaker and punctuated the record at frequent predetermined points by tapping a key which momentarily displaced the timing wave. By the aid of these punctuations we were enabled to identify the words and syllables on the records after development. The areas for computing average power were measured from the mean power curve, while the instantaneous power curve was measured only for its peak values.

Although chosen at random, the speakers used in these tests represent all sections of the United States. Their types range from soprano to bass-baritone, neither extreme type—high soprano and bass—being available; but this assortment is sufficiently representative for our purpose. Extraneous disturbances were to a large extent eliminated by the sound-proofing on the walls and ceiling. Lest the novelty of this situation be a distraction to the speaker, he was allowed to practice and become accustomed to the new condition.

Conclusion

One advantage in having speech data available in terms of its power rather than its amplitude is the fact that in most instruments used for making quantitative speech measurements, the force which operates the meter is proportional to the square of the wave amplitude. Common examples of such instruments are the dynamometer and the ordinary a.c. meters.

To summarize, the power is classified into:

1. Instantaneous power, P_i.
2. Mean power, P_m.
3. Long average power.
4. Short average power.

Stress calibrations are here derived from the maximum values of P_i and P_m (P_I and P_M, respectively) in each syllable, while the use of the total energy of the syllable for calibrating emphasis also shows possibilities. The peak factor is the square root of P_I/P_m and is a useful index of the wave form.

The measuring apparatus—excluding the rectifier and oscillograph—is essentially a good quality speech-transmitting system. In view of the fact that good quality systems are now used commercially as well as in the laboratory the data naturally fall into two classes:

(1) Measurements which characterize the speech solely from the standpoint of the transmitting apparatus;

(2) Estimates or approximations concerning the total power from the voice.

Regarding (1) we note that the divergence of waves causes some frequency distortion which is greater, the nearer the source, and becomes negligible as the distance increases (see the appendix). We should accordingly expect the peak factors to be different at the speaker's lips. The estimates of total power, however, are as close as their importance necessitates.

When the data are applied to a case in which the speaker's distance is other than 9 cm., the required power intensity is found by the law of inverse squares and the pressure by the law of inverse distance.

APPENDIX

Frequency Distortion in Spherical Waves

A spherically diverging sound wave (see H. Lamb: "Dynamical Theory of Sound," page 206) is represented by

$$r\phi = f\,(v_o t - r)$$

where r = radius of the wave front

ϕ = velocity potential

t = time

v_o = velocity of sound

ρ_o = mean density of air

The pressure

$$p = -\rho_o v_o \partial\phi/\partial r$$

$$= \rho_o v_o \left[\frac{1}{r} f'\,(v_o t - r) + \frac{1}{r^2} f(v_o t - r) \right]$$

Let $f(v_o t - r) = \sin \omega \left(t - \frac{r}{v_o} \right)$,

so that

$$p = \frac{\rho_o v_o}{r} \left(\frac{\omega}{v_o} \cos \omega \left(t - \frac{r}{v_o} \right) + \frac{1}{r} \sin \omega \left(t - \frac{r}{v_o} \right) \right).$$

When a wave composed of any number of such components (each having a different pair of values for ω and α) diverges from one radius to a larger one, it not only changes in size, due to the factor $\frac{\rho_o v_o}{r}$ but also in shape, due to the factor $\frac{1}{r}$ in the second term. When r

is large compared with $\frac{c}{\omega}$, this change in shape becomes negligible. In the case of speech, since the source is of finite size the effective radius is somewhat greater that that measured from the speaker's lips, and the wave front is not exactly hemispherical, so the comparison is only qualitative. Nevertheless, a difference in quality of transmitted speech can be detected when the speaker's lips are within 2 cm. of the transmitter diaphragm.

Some Contemporary Advances in Physics IX
The Atom-Model, Second Part[1]

By KARL K. DARROW

G. Recapitulation of the Facts to be Explained

EVERY atom-model that is worthy of notice was designed in view of a certain limited group of facts. That is to say, every valuable atom-model is the invention of somebody who, being acquainted with certain of the ways in which matter behaves, set himself to the devising of atoms of which an assemblage should behave like matter in those ways. Of course, it would be a most wonderful achievement to conceive atoms, of which assemblages should behave like matter in all ways; but this is too exalted an ambition for this day and generation, no man of science bothers with it. Each atom-model of the present is partially valid, not universally; and nobody can rightly appreciate any one of them, unless he knows the facts for which it was designed. I might add that he should also know the relative importance, in the world and in life, of the facts for which it was designed. But this also is too exalted an ambition; we do not know much, if anything, about the relative importance of facts *sub specie æternitatis*, and can hardly refrain from regarding with an especial favour the facts which happen to have been successfully explained. At all events it is clear that every account of an atom-model should be preceded by an independent account of the things it is meant to explain. For the favorite atom of these days, the atom of Rutherford and Bohr, I have provided this preliminary account of the facts in the First Part of the article. Let me give a brief outline of the most important among them, before entering upon the task of constructing an atom-model to reproduce them.

First and foremost, the elements are very definite things; each of the ninety of them is distinguishable from the other eighty-nine, not in one respect only but in many, and in many cases the contrasts are very severe. The atom designed for each of them must therefore have definiteness and fixity and a sharply-marked character.

Next: although the atom must be definite, it must not be absolutely immutable; it must be capable, under stress, of assuming various distinct states or forms or configurations or whatever you choose to call them. This is prescribed by that great and essential fact of the Stationary States, to which so much of the First Part of

[1] Devoted to Bohr's atom model for hydrogen and ionized helium. The models for other atoms, as well as some general considerations, are reserved for the Third Part.

this article was devoted. For an atom, when initially in its normal state and properly stimulated, is able to receive energy in certain definite measurable amounts, and to retain it for a while; and this is tantamount to saying that each atom may exist for a while in one or another of certain states distinct from the normal state, in each of which it possesses a certain distinctive amount of extra energy. Thus a helium atom may receive 19.75 equivalent volts of energy from an impinging electron, no less and (within certain limits) no more; and this is tantamount to saying that a helium atom may exist, not only in its normal state but also transiently in an abnormal state in which its energy is greater by 19.75 equivalent volts than in the normal state. The atom-model for each element must therefore be designed to be definite in each of several distinct and interchangeable states, and not in one only.

The energy-values of some few of these stationary states are determinable directly; but most of them (and they are very numerous) are deduced from spectra. The spectrum of an element is the family of radiations of various frequencies which it emits when it is in the gaseous state. These are commonly ascribed to the individual atoms. The first task of the spectroscopist is to measure these frequencies; his second, to classify them. In certain spectra his task of classification is easy, for there is a natural arrangement of the spectrum lines which "leaps to the eye." This is an arrangement of lines in one or several converging series, like those of which there were photographs of the First Part of the article. Let me represent by

$$\nu_1, \nu_2, \nu_3, \ldots \nu_i, \ldots$$

the frequencies of the consecutive lines of a series, and by ν_{lim} the frequency of the series-limit upon which they converge. Now the frequencies of the various lines may be described by a formula

$$\nu_i = \nu_{lim} - f_i \qquad (1)$$

in which ν_i is expressed as the difference between two *terms*. The term f_i varies from one line to the next; and in some instances this function f_i is algebraically of an extreme simplicity, just the sort of a simple elegance which is apt to suggest that the formula has an inward physical meaning. Also one and the same term may figure in the formulae for lines belonging to different series, a fact which enhances the feeling that the terms are physically "real." Thus the spectroscopist seeks "terms" whereby to classify the lines of a spectrum; and the analysis of a spectrum leads to the measurement of a multitude of terms.

Now multiply both sides of equation (1) by Planck's constant h; it becomes

$$h\nu_i = h\nu_{lim} - hf_i. \tag{2}$$

On the left-hand side we have $h\nu_i$, a quantity of the dimensions of energy. Now there is much reason to believe that when radiant energy streams out from a substance in the form of radiation of frequency ν, it emerges often if not always in parcels or packets or units or *quanta*, each consisting of an amount of energy equal to $h\nu$. Suppose that the radiant energy constituting any line of a series is emitted in quanta such as these; then whenever an atom performs the act of radiating that line, it loses the amount of energy which stands on the left-hand side of Equation (2). The right-hand side represents the same thing, and is itself the difference between two terms which are spectrum-terms multiplied by h; these are themselves the values (reckoned from a suitable zero) of the energy of the atom before and after the radiation occurs, they are the energy-values of the atom in the state before radiating and in the state after radiating. *The spectrum-terms, when multiplied by Planck's constant* h, *are translated into the energy-values of the Stationary States of the atom.* When expressed in proper units, terms are energies and energies are terms. In the decades during which the spectroscopists were analyzing line-spectra, disentangling line-series—by no means a light labor, for the perspicuity of the series shown in the photographs of the First Part is anything but common—and disengaging terms, they were unknowingly recognizing and locating the Stationary States of the atom. Spectrum analysis culminates in the fixation of the Stationary States. This is the greatest of the ideas for which the world is indebted to Bohr, and eventually through him to Planck.

These Stationary States constitute one of the great systems of facts, which the atom-model of Rutherford and Bohr is designed to interpret. Let me formulate the demands which thus are made upon this atom-model. It must have features to account for these facts:

First, that there are such things as Stationary States;

Second, that in passing over in a "transition" from one stationary state to another of which the energy is less by ΔU, the atom releases the energy ΔU in radiation of the one frequency $\Delta U/h$;

Third, that certain transitions do not occur, or occur under abnormal circumstances only, or occur less frequently than others; and

Fourth, that the stationary states of each particular kind of atom have the particular numerical energy-values which they are observed to have.

The first three of these demands are of a general and fundamental nature. If someone were to design an atom-model for these phenomena of the Stationary States and these alone, he would probably begin by imagining an atom which would satisfy these general demands; then he would proceed so to specialize it that it would comply also with the fourth. It might have been well, had this happened; the course of history was otherwise. The atom-model of Rutherford was designed originally to interpret phenomena of quite another field, and then Bohr modified it by violence to satisfy the fourth of the foregoing demands.

Of the facts which Rutherford devised his atom-model to interpret, the cardinal one is that the atom contains electrons. The best evidence for this fact is, that electrons can be extracted from atoms.[2] One can even measure the amount of energy required to extract an electron from an atom — in other words, the difference between the energy of an atom in its normal state, and the energy of the same atom in its "ionized" state.[3] This has a direct bearing on the phenomena of the Stationary States; for the spectrum-terms, when they are multiplied by Planck's constant h, yield the energy-values of the corresponding Stationary States, reckoned from the energy-value of the ionized state as zero of energy.

Granted that the atom contains electrons: it must contain positive electricity also, to compensate their negative charge. Now it is easy to imagine the positive electricity so arranged, that the electrons can be fitted into various places within and around it, and remain in equilibrium[4]; it is possible to imagine that the positive electricity acts upon the electrons with a force which is compounded of the familiar inverse-square attraction and a particular sort of a repulsion, so adjusted that the electrons will remain in equilibrium in various positions. It seems as though the Stationary States might be interpreted in this fashion, and several attempts have in fact been made; but they are discouraged by the experiments of Rutherford and his followers on the deflections of alpha-particles and electrons which pass through atoms. For these deflections occur exactly as if the positive electricity were concentrated at a point or "nucleus," and an inverse-square electric field prevailed in the region between this nucleus

[2] This is not quite a proof of the fact. As Aston cleverly remarked, when a pistol is fired, smoke and a bullet come out of it; we are quite justified in inferring that the bullet was originally within the pistol, but not the smoke!

[3] This energy, which I called the energy of the "state of the ionized atom" in the First Part, is truly the energy of the system composed of the atom minus its electron, and the free electron.

[4] Although not in stable equilibrium.

and the electrons.[5] They may be compatible with other atom-models; it is certainly incumbent upon the designer of any other to prove that they are compatible with his. Furthermore these deflections indicate that the positive charge on the nucleus of the atom is just sufficient to compensate the negative charges of a number N of electrons, equal to the "atomic number" Z which is the cardinal number defining the position of the element in the Periodic Table of the Elements. This confirmation of the splendid idea of van den Broek and Moseley is so delightful and so precious, that anyone would hesitate long before rejecting the atom-model whereby it is deduced from Rutherford's experiments.

Yet this *nuclear atom-model* cannot be accepted, without being instantly modified. A system consisting of a positively-charged nucleus and electrons surrounding it, all acting upon one another with inverse-square forces of attraction between nucleus and electrons and repulsion between one electron and another, is not a stable system; it is a suicidal system, doomed to quick and permanent collapse. If the electrons were initially standing still, they would fall into the nucleus; if the electrons were initially swinging in orbits about the nucleus like planets around the sun, they would steadily radiate their energy into space—not in radiation of one single frequency either, but in a mixture of all possible frequencies—and would wind their ways spirally into the nucleus. Therefore, the nuclear atom-model must be altered; for instance, by adding a proviso, that the electrons shall stand still, and shall not be sucked into the nucleus; or a proviso, that the electrons shall revolve in closed orbits planetwise, without radiating any of their energy [6], and without gliding by a spiral path into the nucleus.

Suppose then that we decide to make one or the other of these provisos, in order to save the interpretation of Rutherford's experiments. Could we then so shape the proviso, that it would satisfy the four demands which I described as being made upon the atom-

[5] Apart from such deviations in the immediate neighborhood of the nucleus as the most delicate experiments of this sort reveal; which cannot be supposed to extend to the region where the electrons are.

[6] To indicate how much this neglect of the radiation from the revolving electron amounts to, I cite the results of a calculation given by Wien in his lecture *Ueber Elektronen*, and doubtless elsewhere. Imagine an electron distant by ten Angstrom units from a hydrogen nucleus, and moving with such a velocity that, but for the radiation, it would revolve in a circle about the nucleus. In a single circuit, it should radiate about one ten-millionth part of the kinetic energy it initially possesses. Hence the single circuit will differ very little indeed from a perfect circle; and in this sense, the radiation is truly negligible. But the single circuit is described in less than 10^{-20} second; hence, in any time-interval long enough to be measured by the most delicate of physical apparatus, the dissipation of energy by radiation is far too great to be neglected with impunity.

model by the facts of the Stationary States? Could we for instance so shape the first proviso, *could we choose such locations for the electrons assumed stationary*, that the sodium atom (for instance) would display only those energy-values which the spectrum of sodium allows for its Stationary States, and no others?

Undoubtedly we could. The sodium atom is supposed to consist of eleven electrons surrounding a nucleus of charge $+11e$. If the electrons were all stationary in assigned positions about the nucleus, we could calculate the energy of the arrangement. The energy-values of the various Stationary States being known, it would not be difficult to find, for each one of the Stationary States, at least one arrangement of the eleven electrons identical with it as to energy-value. Having done this, we could lay it down as a law that the electrons shall stand still in each and any one of these arrangements; but not in any other arrangement whatsoever.

But would this be an explanation of the Stationary States? Not, I think, in any significant sense of that valuable word. It could justly be designated as an explanation, as a theory, only if the various arrangements so prescribed for the various Stationary States should turn out to be interrelated according to some law—to be governed by some unifying principle—to display some intrinsic quality of simplicity and elegance and beauty, distinguishing them from all the other and rejected arrangements. This has not been achieved.

Let me now take up the other of the two suggestions which were made above. Suppose that we accepted the nuclear atom-model, with the proviso that the electrons should revolve in closed orbits planetwise, without radiating any of their energy, and without gliding by a spiral path into the nucleus. Could we so shape this second proviso, *could we choose such orbits for the electrons assumed revolving without loss of energy*, that the sodium atom or the hydrogen atom (for instance) would display only those energy-values which the spectrum of sodium or the spectrum of hydrogen prescribes for the Stationary States, and no others?

Again, there is no doubt that we could; but the value of the achievement, again, would depend on whether or not the orbits which we thus selected were interrelated according to some law, or governed by some unifying principle, or distinguished from all the other orbits by something seemingly fundamental. Consider Rutherford's model for the hydrogen atom, which consists of a nucleus and an electron. If we adopt the proviso which was just set forth, and suppose that the electron may revolve around the nucleus in circular orbits without radiating any of its energy, then we can select particular circular orbits, such that when the electron is revolving in one or another of these, the energy of the atom shall have one or another of the values prescribed by the Stationary States. If we arbitrarily say that the electron can revolve only in one or another of these orbits, then we have an atom-model competent to interpret the Stationary States of the hydrogen atom. But is there anything distinctive about these selected orbits, anything peculiar, anything which marks them out and sets them apart from the other, from the discarded orbits? Have they any feature in common, apart from being necessary to give the observed energy-values of the Stationary States?

It is hardly possible to lay too strong an emphasis upon this requirement; the value of the contemporary atom-model depends upon satisfying it. Let me put the matter another way. From the moment that we imagine that the electrons within the atom are cruising around the nucleus in orbits without radiating energy and without dropping into the nucleus, we are sacrificing the unity and the coherence of the classical theory of electricity. So grave an action is not to be undertaken lightly nor with indifference; it were foolish to make such a sacrifice without recompense; and there is no recompense to be found in merely proving that especial orbits can be so selected as to copy the energy-values of the Stationary States. If one is going to deviate from the rules of the classical theory of electricity, one must deviate by rule. If one is going to disrupt the system which prevails in one great department of theoretical physics, one must systematize another department in exchange. If one proposes to violate some of the principles of modern physics, by asserting that electrons can travel in certain orbits without radiating, he must reconcile the congregation of physicists to his sacrilege by proving that the selected motions are themselves governed by a principle, as imposing as those he lacerated. If the innovator cannot show that his innovations are systematic, he is not likely to prosper; but if his innovations are derived from a principle, it may supersede those which he contradicted.

To discover such a principle is the ambition of, probably, half of the theoretical physicists who are active today.

There are other general statements which might be made at this point; but they will be more intelligible, and so will the foregoing paragraphs be, after I have given an illustration. For this purpose I will describe two models of the hydrogen atom, each of them consisting of a nucleus and a single electron, each capable of being so constrained that its energy-values will copy those of the Stationary States of hydrogen. With one of these, however, the description can be carried no farther. With the other, I shall show—following Bohr—that the orbits in which the electron is constrained to revolve have certain peculiar features, distinguishing them above all other orbits; and these distinctive features may be consequences of the desired and still hidden principle.

II. Features of the Necessary Orbits of the Hydrogen Atom (Quantization)

Hydrogen being the first element in the periodic table, Rutherford's atom-model for it consists of a nucleus and one electron. The electron bears (or *is*) a negative charge amounting to $-e$ or $-4.774.10^{-10}$ electrostatic units, and its mass is approximately 9.10^{-28} grammes. The nucleus bears a positive charge amounting to $+e$, and its mass is about 1,840 times as great as that of the electron.

The stationary states of the hydrogen atom possess the energy-values $-Rh$, $-Rh/4$, $Rh/9$, $-Rh/16$, $-Rh/25$, and so on; in general, the values $-Rh/n^2$ $(n=1,2,3\ldots.)$. The constant[5] R is equal to $3.29.10^{15}$; the constant h is Planck's constant $6.56.10^{-27}$ erg. sec.

Rutherford's atom-model for the hydrogen atom must now be so modified, that it will admit the energy-values just specified, and no others.

I will begin by doing something which amounts to setting up a straw man, to be knocked down immediately, but not, I hope, before he does us some service. Let us suppose that, in spite of all the laws of dynamics, the electron may stand still at a distance r from the nucleus, without starting towards and falling into it. With the electron in such a position, the energy of the atom is $-e^2/r$. This is an energy-value referred, like all energy-values, to a particular zero; in this case, the zero-value of energy corresponds to the condition in which the electron is infinitely far away from the nucleus. We recognize at once the "state of the ionized atom," to which the energy-values of the Stationary States as given by the spectrum-terms are automatically referred. This quantity $-e^2/r$ must be permitted to assume the successive energy-values of the successive Stationary States, and no others; we must have

$$
\begin{aligned}
-e^2/r &= -Rh && \text{for the first (or normal) stationary state}\\
-e^2/r &= -Rh/4 && \text{for the second stationary state} \qquad (3)\\
-e^2/r &= -Rh/9 && \text{for the third stationary state; and so forth.}
\end{aligned}
$$

[5] I deviate here from the more frequent usage of defining R from the equation

$$\frac{1}{\lambda} = R\left(\frac{1}{m^2} - \frac{1}{n^2}\right)$$

for the reciprocals of the wavelengths of the various lines of hydrogen; in which equation $R = 109677.69$ by measurements of tremendous accuracy, and is to be multiplied by c to get what I have called R.

Now each of these equations defines a value of r; we have

$$r = e^2/Rh \text{ for the normal state}$$
$$r = 4e^2/Rh \text{ for the second stationary state} \quad (4)$$
$$r = 9e^2/Rh \text{ for the third stationary state; and so on.}$$

Each of these values of r represents the distance at which the electron must stand from the nucleus, that the atom may have the energy-value of the corresponding stationary state. If we say that the electron may stand still at and only at the distances given by

$$r = e^2/Rh,\ 4e^2/Rh,\ 9e^2/Rh, \ldots\ldots\ldots\ldots, \quad (5)$$

we thus define an atom-model interpreting the Stationary States. It is scarcely an atom-model to be recommended, and I certainly am not taking the responsibility of recommending it. Nevertheless the reader had best beware of picking out the obvious objections to it, and condemning it because of them. For if he objects that I have given no reason why the electron should stand still at all, nor why it should stand still in these and only in these positions, nor why it should cause radiation of a peculiar and well-defined frequency when it passes from one of these positions to another if he makes these objections, I can retort that the atom-model favored by Bohr himself suffers from every one of these deficiencies. In fact, the only defects peculiar to this "atom-model of the stationary electron" appear to be two. The first is, that the distances specified by (5) do not have distinctive features such as I shall presently show for the orbits specified for the "atom-model of the revolving electron"; and this defect, as I have tried to emphasize, is a grave one. The second is, that an atom in which the charges are stationary is not *ipso facto* magnetic, whereas an atom with revolving electrons is.[7]

Following Bohr, and practically all the other physicists of today, we now assume that the electron revolves planetwise around the nucleus describing a closed orbit and radiating none of its energy as it revolves. A planet revolves in an elliptical orbit; this elliptical orbit may be a circle, or it may not be; but for the present paragraph we will think of the circles only. Let us suppose, then, that the electron may revolve in a circle about the nucleus, without radiating its energy and spiralling into the nucleus. Designate the radius of the circle by r. With the electron revolving in a circle of radius r, the energy of the atom is $-e^2/2r$. This value is obtained by adding together the potential energy of the atom, which is $-e^2/r$ just as it

[7] If any reader can abolish these defects, a multitude of chemists will be glad to hear from him. Chemists want atom-models with stationary electrons.

was when we supposed the electron to be standing still, and the kinetic energy of the electron, which is $\frac{1}{2}mv^2$. In this last expression, v stands for the speed of the electron in its orbit; now, mv^2/r is the "centrifugal force" acting upon the electron, which is equal (and opposite) to the attraction exercised by the nucleus upon the electron, which is e^2/r^2; so that $\frac{1}{2}mv^2$ is equal to $+e^2/2r$, and the total energy of the atom has the value $-e^2/2r$. As before, this is the energy-value referred to the state of the ionized atom.

This quantity $-e^2/2r$ must be permitted to assume the successive energy-values of the successive Stationary States, and no others; we must have

$$-e^2/2r = -Rh/n^2 \quad (n = 1, 2, 3, 4, \ldots\ldots) \tag{6}$$

Each of these equations defines a value of r, as follows:

$$r = n^2e^2/2Rh \quad (n = 1, 2, 3, 4, \ldots\ldots) \tag{7}$$

If we say that the electron may revolve in and only in such circles as have the radii given by the equations (7), we thus define an atom-model interpreting the Stationary States. Is this atom-model superior to the tentative one which was described just before it? Not in any way which has yet been brought to notice. No reason is given why the electron should revolve in a circle instead of spiralling into the nucleus, nor why it should revolve in these and only in these circles, nor why it should cause radiation of a peculiar frequency to be emitted when it passes from one of these circles into another. All of the objections which I suggested, a few paragraphs above, that the reader might raise against the then-mentioned atom-model with the stationary electron, may equally well be raised against this atom-model with the revolving electron. Why then should we attach greater importance to this one than to that? Partly, as I said, because this atom possesses intrinsic magnetic properties, while to the other one magnetic qualities would have to be ascribed by an additional assumption; but chiefly because Bohr discovered certain distinctive features of the circular orbits defined by (7), which set them apart from all others. These we now examine.

To understand the first of these features, it is necessary to consider the angular momentum of the atom. Sooner or later we shall have to make a slight alteration in the reasoning indicated in the last paragraphs; it may as well be made now even though it is not yet necessary. Heretofore I have tacitly assumed that the nucleus stands still while the electron revolves around it. As a matter of fact, if the atom may be represented as a solar system in miniature, the nucleus and the electron both revolve about their common centre of mass in ellipses—we will think, as before, only of circles (Figure 1). The radii a and A of the circular orbits of the nucleus and the electron, being the respective distances of the particles from their centre of mass, stand in the reciprocal ratio of the masses M of the nucleus and m of the electron; and as they describe their orbits in the same period (since the centre

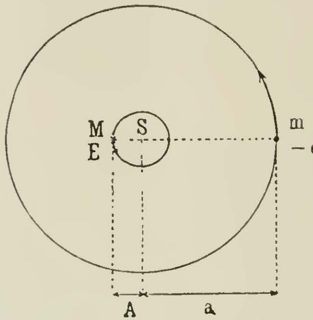


Fig. 1—Diagram to illustrate how the electron and the nucleus revolve around their common centre of mass in synchronous orbits

of gravity is at rest and always between them) their speeds v and V stand in the same ratio:

$$a/A = v/V = M/m. \tag{8}$$

I introduce the symbol μ to denote the equal quantities

$$\frac{M}{M+m} = \frac{a}{a+A} = \frac{v}{v+V}. \tag{9}$$

The potential energy of the atom, reckoned as always from the state in which the nucleus and the electron are infinitely far apart, is obviously $-e^2/(a+A) = -e^2\mu/a$. The kinetic energy of the atom is the sum of the portion $\frac{1}{2}mv^2$ belonging to the electron and the portion $\frac{1}{2}MV^2$, belonging to the nucleus. I point out that the "centrifugal force" acting upon the electron is mv^2/a, and that acting upon the nucleus is MV^2/A, and each of these separately must be equal to the reciprocal attraction $e^2/(a+A)^2$ of nucleus and electron; and I leave it to the reader to show by means of these equalities that the kinetic energy amounts to $\frac{1}{2}e^2\mu/a$. The total energy of the atom is therefore equal to $-\frac{1}{2}e^2\mu/a$, and this is the quantity to be equated to the observed energy-values of the stationary states; equation (6) is replaced by

$$-e^2\mu/2a = -Rh/n^2. \qquad (10)$$

The angular momentum of the electron is mva; the angular momentum of the nucleus is MVA; the angular momentum of the atom, for which I use the symbol p, is the sum of these:

$$p = mva + MVA = mva/\mu. \qquad (11)$$

I leave it again to the reader to use the foregoing statements to arrive at the expression

$$p = e\sqrt{ma} \qquad (12)$$

and by combining (12) and (10), at the expression

$$p_n = ne^2\sqrt{m\mu/2Rh} \qquad (13)$$

for the value p_n of the angular momentum of the atom, or rather of our atom-model, in its nth stationary state.

Thus the values of the angular momentum of the atom-model, in the various states in which it has the prescribed energy-values $-Rh$, $-Rh/4$, and so forth, increase from the first of these states onward in the ratios 1:2:3:4 . . . They are the consecutive integer multiples of a fundamental quantity, the quantity

$$p_1 = e^2\sqrt{m\mu/2Rh}. \qquad (14)$$

Now it happens that this fundamental quantity is equal, within the limits of experimental error, to $h/2\pi$—to $1/2\pi$ times that same constant h which has already figured in this discussion:

$$p_1 = h/2\pi; \; p_n = nh/2\pi. \qquad (15)$$

This occurs because the value of R is equal, within experimental error, to the combination of m, e, and h on the right of this equation:

$$R = 2\pi^2\mu me^4/h^3. \qquad (16)$$

The atom-model which I have been describing at some length could therefore be described in a few words by saying that *the electron is permitted to revolve only in certain circular orbits, determined by the condition that the angular momentum of the atom shall be equal to an integer multiple of* h/2π. This condition is in fact sufficient to impose the values given for the radii of the circular orbits in equations (10) which values in turn entail the desired energy-values for the stationary states. The reader can easily prove this by working backward through the train of equations; and indeed this is the manner in which the Bohr atom-model is usually presented, so as to arrive finally at the agreement between "theory" and experiment which is expressed in equation (16), and is a most striking climax to the whole exposition. By working through the train of equations in the inverse sense, I have considerably mitigated the effect of the climax; and this procedure seems hardly fair to the author of the theory, but it is not without its merits, for it enables us to see the exact role of equation (15) more clearly than the commoner procedure.

The situation now is this. It is possible to construct, out of a nucleus and an electron, an atom-model possessing stationary states of the energy-values displayed by the hydrogen atom, provided that we assume that the electron may revolve only in circular orbits for which the angular momentum of the atom is an integer multiple of $h/2\pi$. There is no known reason why an electron should do a thing like this, there is good reason to suppose that it cannot do anything of the sort, for if it started out to revolve in a circular orbit it would radiate its energy and descend spirally into the nucleus. If nevertheless we assert that the electron does just this sort of thing, we have nothing with which to support the assertion, nothing extrinsic by which to render it plausible; it must stand on its own merits as an independent principle.

These merits, had we no data other than the energy-values of stationary states catalogued in equation (6), would probably be regarded as scanty. After all, the agreement between the constant p_1 and the quantity $h/2\pi$ might be fortuitous. But there are other stationary states of the hydrogen atom, beyond those listed in (6). For instance there are the stationary states which are evoked by a strong electric field acting upon hydrogen, and there are the stationary states which are called into being by a magnetic field applied to hydrogen, as I related in earlier sections of this article. There is also the fact, that at least one of what I have been calling the stationary states of hydrogen is not a single stationary state at all; there are two states of which the energy-values lie exceedingly close together and to the value $-Rh/4$, so close that nearly all experiments fail to discriminate them. And there is the great multitude of stationary states exhibited by other elements than hydrogen; but we will not think about these for the time being.

Now the situation is transformed into this. Consider all these additional stationary states, exhibited by the hydrogen atom under unusual or even under usual circumstances. Is it possible to trace, for each one of them, an orbit for the electron, such that while the electron is describing that orbit, the energy of the atom possesses just the value appropriate to that Stationary State? And granting that this is possible and accomplished; can it be shown that these additional orbits are distinguished by some feature resembling that feature of the circular orbits which is described by equation (15)? Our condition laid upon the circular orbits, that in each of them the angular momentum of the electron is an integer multiple of $h/2\pi$—this condition valid for the limited case, can it be generalized into a condition governing the Stationary States of the hydrogen atom under all circumstances? Can orbits be described which account for all of the Stationary States of hydrogen under all circumstances, and which are determined by a general condition of which the condition set forth in equation (15) is one particular aspect? If so, that general condition might well be such a Principle as the one towards which, as it was said in the last section, so many physicists aspire. Thus the test to which this condition laid upon the angular momentum must be submitted is this: *can it be generalized?*

Before trying to generalize it let us examine some other distinctive features of the circular orbits defined in (7)—I will call them henceforth the "permissible" circular orbits, but we should remember that perhaps it is only ourselves who are "permitting" them and forbidding the others, and not Nature at all. Let us calculate the integral I of the doubled kinetic energy $2K$ of the atom over a complete revolution of the electron (and nucleus):

$$I = \int_0^T 2K dt. \tag{17}$$

It is easy in this case, for K is constant in time, so that $I = 2KT$. Now K is equal to $\frac{1}{2}mv^2/\mu$, and T is equal to $2\pi a/v = 2\pi^2 ma^2/\mu K$; which expression the reader may reduce, by means of that equation $K = \frac{1}{2}e^2\mu/a$ which he was invited to derive, to

$$T = \pi e^2 \sqrt{m\mu/2K^3} \tag{18}$$

multiplying which by K, and using equation (10), we have

$$I = 2\pi n \cdot e^2 \sqrt{m\mu/Rh}. \tag{19}$$

The reader will recognize the factor which appeared in (11) and was there stated to be numerically equal, within the error of observation, to $h/2\pi$.

Therefore this atom-model could also be described by saying that *the electron is permitted to revolve only in certain circular orbits determined by the condition that* I *shall be equal to an integer multiple of* h.

For future use I interpolate the remark that the factor n is called the *total* or *principal quantum number*; in German, *Hauptquantenzahl.*

The reader will think that this is not a new condition, but only a futile way of re-stating the condition laid upon the angular momentum. So it might be, in this case. But when we come to the more complex cases, we shall find that the two conditions diverge from one another. *Which of the two can be generalized, if either?* Only experience can show.

I will describe one more distinctive feature of the permissible orbits; it may seem more impressive than either of the others.

We have seen that the frequency of the radiation emitted, when the hydrogen atom passes from one stationary state to another—say from the state of energy $-Rh/n'^2$ to that of energy $-Rh/n''^2$—is

$$\nu=\frac{R}{n''^2}-\frac{R}{n'^2}$$

which may be written

$$\nu=\frac{R}{n'^2n''^2}(n'-n'')(n'+n''). \tag{20}$$

Suppose that $n'-n''=1$, that is, that the transition occurs between two adjacent stationary states of the atom; and let n' and n'' increase indefinitely. In the limit we shall have

$$Lim\ \nu=\frac{2R}{n'^3}. \tag{21}$$

Accepting the atom-model with the electron revolving in a circular orbit, we take from (18) the value for the period of the revolution, substitute for K by the aid of (10), and arrive at this expression for the frequency of the revolution:

$$\omega'=v/2\pi r=\sqrt{8R^3h^3}/2\pi n'^3e^2\sqrt{m\mu} \tag{22}$$

Comparing this expression for ω' with the expression for $Lim\ \nu$ in (21), we see that they are identical, if

$$R=2\pi^2m\mu e^4/h^3$$

and this will be recognized as being that very value of R which was given in equation (13), as the value established by experiment. Thus the experimental value of R is such that

$$Lim\ \omega=Lim\ \nu. \tag{23}$$

In this equation the symbol ω stands for the frequency of revolution of the electron in its orbit, when the energy of the atom is $-Rh/n^2$. It therefore stands for the frequency of the radiation which the atom would be expected to emit; for an electrical charge performing a periodic motion should, according to the fundamental doctrines of the electromagnetic theory, be the origin of a stream of radiation with period equal to its own. The symbol ν stands for the frequency of the radiation which the atom does emit in passing between two adjacent Stationary States. According to (19), this actual frequency is more nearly equal to the expected frequency, the more remote these two adjacent Stationary States are from the normal State; and in the limit, actual frequency and expected frequency merge into one. The numerical value of the constant R is just such as to bring about this relation.

Here again we have a curious numerical agreement which, like the other correlated fact that the angular momentum of the electron in the nth orbit is $nh/2\pi$, may by itself be merely a coincidence; but this one has a much greater inherent appeal. We have relinquished the expectation that the electron, cruising around the nucleus in a cyclic path, will send forth radiation of the frequency of its own revolutions, as every inference from the laws of electricity indicates that it should; but here is a case—even if it is only a limiting case—in which the frequency emitted from the atom agrees with the one which we should expect. Generally there is discord; but in the limiting case there is consonance. Does this not suggest that the desired Principle may be one which in a limiting case merges with the classical theory of electricity—possibly, indeed, nothing less than the foundation of a general theory of electricity, of which the classical theory expresses only a special case?

Let us review our situation.

Having supposed for hydrogen an atom-model consisting of a nucleus and an electron;

Having supposed that these revolve around their common centre of mass according to the laws of dynamics, but without spending any energy in radiation;

Having supposed in particular that they revolve only in circular orbits, and only in such circular orbits as yield for the atom-model the energy-values $-Rh/n^2$ measured by experiments upon the Stationary States;

Having traced these "permissible" circular orbits,

We have found that they are distinguished from all the other circular orbits by at least three peculiar features (viz., the features expressed by the equations $p = nh/2\pi$, and $I = nh$, and $Lim\ \omega = Lim\ \nu$).

We do not know that there is any revolving electron at all. We know only that if all our suppositions be correct, the consequences expressed by these three equations are correct also. Are these consequences impressive enough to prove the suppositions true?

The answer to this question depends on our degree of success, or rather on the degree of success attained by Sommerfeld and Bohr and their followers, in generalizing these equations to other and more complex cases. Usually the process of generalizing will involve difficult labours of orbit-tracing. But it is possible to make a significant comparison between the spectra of hydrogen and of ionized helium, without additional studies of orbits.

I. Relations Between the Spectrum of Hydrogen and the Spectrum of Ionized Helium

To make trial of the validity of the foregoing ideas about the origin of the hydrogen spectrum, one naturally applies them to whatever other spectra may reasonably be ascribed to an atom consisting of a nucleus and a single electron. As according to the view adopted in this article the atom of the nth element in the Periodic Table consists of a nucleus and n electrons, the only way to produce such a spectrum is to produce a sufficient number of atoms of some element or other, each atom lacking all but one of its electrons; helium atoms deprived each of one electron or "once-ionized," lithium atoms deprived each of two or "twice-ionized," beryllium atoms deprived each of three electrons, or in general atoms of the nth element of the Periodic Table divested each of $(n-1)$ electrons. This we should expect to require violent electrical or thermal stimulation of the vapor of the element, more violent the more electrons have to be removed. Hence it is not surprising that the spectrum of once-ionized helium is the easiest of these spectra to produce; but it is more than a little strange that this is not merely the easiest but the only spectrum of this kind which has ever been obtained. Even the spectrum of twice-ionized lithium has not been generated, in spite of efforts quite commensurate with the value it would have.[8] The spectrum of once-ionized helium remains the only companion of the spectrum of hydrogen; these are the only two known spectra which are ascribed to atoms consisting of a nucleus and a single electron.

We have seen that if we imagine that the electron of the hydrogen atom can revolve, without spending energy by radiation, in and only in those circular orbits for which the angular momentum of the atom is equal to $h/2\pi$, $2h/2\pi$, $3h/2\pi$, ... $nh/2\pi$, ... , then the energy of the atom-model can assume only the values $-Rh$, $-Rh/4$,

[8] Consult for instance the article by Angerer, *ZS. f. Physik*, 18, pp. 113 ff.

$-Rh/9, \ldots. -Rh/n^2$, which are the energy values for the observed stationary states of hydrogen. If this is not an accidental coincidence, then by imagining that the electron of the ionized helium atom likewise can revolve only in orbits for which the angular momentum of the atom is some integer multiple of $h/2\pi$, and by calculating the corresponding energy-values for the atom-model, we should arrive at the energy-values of the observed stationary states of ionized helium. Now the charge on the nucleus of the helium atom is $2e$, twice the charge of the hydrogen nucleus; the force which it exerts on an electron at distance r is $2e^2/r^2$, instead of e^2/r^2. If the reader will work through the equations of Section H, making this alteration wherever appropriate, he will find for the energy-values of the stationary states the sequence

$$-4Rh, -4Rh/9, -4Rh/16, \ldots -4Rh/n^2, \ldots$$

in which

$$R = \frac{2\pi^2\mu m e^4}{h^3} \qquad (25)$$

as heretofore. The quantity μ will be different from what it was for hydrogen; but the difference will be very slight. Therefore if the condition that the electron may revolve about the nucleus only in circular orbits for which the angular momentum of the atom is $nh/2\pi$ is an essential condition, and governs the atoms of hydrogen and ionized helium alike, the stationary states of ionized helium correspond one-to-one with those of hydrogen, but with energy-values almost exactly four times as great. So also with the lines of the spectrum; to each line of the hydrogen spectrum should correspond a line of fourfold frequency in the ionized-helium spectrum; the spectrum of ionized helium should be the spectrum of hydrogen on a quadrupled frequency-scale.

This conclusion is verified. The historical sequence of observations and theories is rather interesting. Certain lines of ionized helium were earliest observed in stars; their simple numerical relations with hydrogen lines being noticed, they were naturally ascribed to hydrogen, and when they were generated in mixtures of hydrogen and helium within a laboratory they were still attributed to the first-named of these gases. Bohr in his first published paper reasoned in the manner I have followed in this section, and inferred that these lines really belonged to helium; which was shortly afterwards verified by seeking and finding them in the spectrum of helium made as pure as possible. A number of additional lines of the spectrum have since been found, although the lines corresponding to transitions into the normal state (the state of energy $-4Rh$) are so far out in the ultraviolet region of the spectrum that no one has yet succeeded in detecting them.

We will now take account of the fact that the numerical values of the constant R calculated for hydrogen (equation 16) and for ionized helium (equation 25) are not quite the same; they are in fact proportional to μ, the quantity which determines the motion of the nucleus, and which varies from one atom to another. In particular

$$R_{He}/R_H = \mu_{He}/\mu_H = (1 + m/M_H)/(1 + m/M_{He}) \qquad (26)$$

in which the symbols m, M_H, M_{He} denote the masses respectively of the electron, the hydrogen nucleus and the helium nucleus, which stand to one another as .000542; 1.000:3.968. Consequently the right-hand member of equation (25) is equal to 1.000403, and the ratio of the frequencies of corresponding lines in the spectra of ionized helium and of hydrogen is

$$4\ R_{He}/R_H \text{ calculated} = 4.001612 \qquad (27)$$

The values of R_{He} and R_H deduced from frequency measurements yield the ratio

$$4\ R_{He}/R_H \text{ observed} = 4.0016212 \qquad (28)$$

The very-exactly-known observed value lies well within the margin of uncertainty of the calculated value. The calculated value of the ratio depends on otherwise-made measurements of the mutual ratios of the three masses (those of the electron, the hydrogen nucleus, the helium nucleus). These otherwise-made measurements are not of the grade of precision claimed for the measurements of $4\ R_{He}/R_H$ by the observations on the spectra. Hence if we combine the observed value of the ratio $4R_{He}/R_H$ with (for instance) the ratio M_{He}/M_H derived from density-measurements upon the two gases, we can calculate a value for the ratio M_H/m ostensibly much more precise than the amount ascertained by direct measurement. This value is

$$M_H/m = 1847. \qquad (29)$$

Let me state briefly what the numerical agreement between the "calculated" and "observed" values of $4R_{He}/R_H$ specifies. It is a test of this set of assumptions; the hydrogen atom and the ionized helium atom may each be represented by a single electron and a nucleus of charge $+e$ in one case and $+2e$ in the other; each stationary state corresponds to a certain circular orbit of the electron; *the Angular Momenta of the two atoms are identical when they are in corresponding stationary states*. As a test, it is favorable. It does not involve the relation between angular momenta and integer multiples of $h/2\pi$ which was stressed in the foregoing section. It is independent of that relation, and may fairly be considered as the second numerical agreement offered by this atom-model, if that relation be considered the first. The idea is due to Sommerfeld; the data whereby the test was made were obtained by Paschen, as a by-product of the work cited in footnote 12.

Although the statements in the foregoing paragraphs are literally true, they do not prove that the condition *Angular Momentum* $= nh/2\pi$ is the distinctive feature *par excellence* of the permissible circular orbits. The result would have been exactly the same if I had defined the stationary states of the ionized helium atom as those for which $I = nh$ or as those for which $Lim\ \omega = Lim\ \nu$.

J. Tracing of Orbits

We must now seek for opportunities to make and test generalizations of the notions about the hydrogen atom explained in section H.

I began by saying that the electron should be supposed to revolve in the inverse-square electrostatic field of the nucleus, according to the laws of dynamics, without spending energy in radiation; and continued by saying that I should speak of circular orbits only. Now the laws of dynamics prescribe elliptical orbits, of which the circular orbits are but special cases. In fact, for each one of the sequence of energy-values $-Rh/n^2$ corresponding to the sequence of Stationary States, there is an infinity of elliptical orbits possessing that energy-value, of which the circle of radius specified by equation (7) is only one. Suppose we should inquire what, if any, are the distinctive features of these elliptical orbits which set them apart from all others?

Again: when radiating hydrogen is exposed to a strong electric field, new stationary states appear, and their energy-values are known. The orbit of an electron, in a field compounded of an inverse-square central field and another field uniform in magnitude and direction, is no longer a circle nor even an ellipse nor even a closed orbit (except in special cases). Could the orbits having energy-values equal to those of the stationary states be identified and traced, and could distinctive features be found which mark them out from among all the others?

Again: when radiating hydrogen is exposed to a strong magnetic field, new stationary states appear, and their energy-values are known. Could the orbit of an electron in a field compounded of an inverse-square central electric field and an uniform magnetic field be traced? and could the orbits having energy-values equal to those of the stationary states be identified? and could peculiar features be found which mark them out from all the others?

Or conversely: is it possible to make "trial" generalizations of one or another of the conditions $p=nh/2\pi$ and $I=nh$ and $Lim\ \omega=Lim\ \nu$? to invent features for the more complex orbits, which sound like reasonable generalizations of these features of the simplest ones? and, having done so, to trace the orbits exhibiting these "trial" features, determine their energy-values, and compare these with the observed energy-values of the stationary states?

Whichever of these two ways is employed to attack the problem, it is necessary to trace orbits more complex, and usually in more complex fields, than the circular orbits imagined for the hydrogen atom. This problem of tracing orbits is the fundamental problem of Celestial Mechanics—the oldest and the most richly developed department of mathematical physics, which in its two centuries and more of history has developed a language and a system of procedures all its own. It is chiefly on that account that many of the recent articles on the atom-model of Bohr are so excessively difficult for any physicist, unless he is of the few who practiced the arts of theoretical astronomy diligently and for a long time before passing over into the field of physics.

In this section I shall quote the equations for the motion of a particle in an ellipse under the influence of an inverse-square central field, and give the derivation with all necessary detail. For the other relevant cases—motion of an electron in a central electric field upon which an uniform electric field, or an uniform magnetic field, or a small central field varying according to some other law of distance than the inverse square, is superposed—I shall give only some of the results, without even attempting the derivation. I shall make no allowance for the motion of the nucleus; the electron will be supposed to revolve around the nucleus considered as fixed. The very small correction required to take account of the motion of the nucleus can easily be applied by the reader, if he so desire. The principal disadvantage involved in neglecting it is, that one too easily thinks of the angular momentum of the electron in its orbit as belonging to the electron alone, whereas it is really the angular momentum of the atom-model. I shall also put E for the charge on the nucleus; E will be equal to e for the hydrogen and to $2e$ for the ionized-helium atom-model, no other cases matter for the time being.[3]

II. Motion of an Electron in an Inverse-Square Central Field

Most people recognize the equation of the ellipse most easily in the form

$$x^2/a^2+y^2/b^2=1$$

in a coordinate-system of which the origin is at the centre of the ellipse, the x-axis and the y-axis parallel respectively to the major and the minor axes of the ellipse.

The symbol a and b denote the semi-major and semi-minor axes of the ellipse; they are related by

$$b^2=a^2(1-\epsilon^2) \tag{30}$$

in which ϵ stands for the "eccentricity" of the ellipse. The foci of the ellipse lie on the major axis at distances $a\epsilon$ to either side of its centre. Transferring the origin to one focus, say the focus at $x=+a\epsilon$, and using coordinate-axes parallel to the former ones, we have

$$(\xi+a\epsilon)^2/a^2+y^2/b^2=1$$

Transforming coordinates again, this time into polar coordinates r and ϕ with the origin at the focus of the ellipse and the direction $\phi=O$ pointing along the x-axis, by means of the substitutions

$$\xi=r\cos\phi \qquad y=r\sin\phi$$

we arrive after somewhat tedious but not difficult algebra[4] at the equation for the ellipse in the form in which we shall use it

$$r=\frac{a(1-\epsilon^2)}{1+\epsilon\cos\phi}$$

and at the derivative thereof

$$\left(\frac{dr}{d\phi}\right)^2=\frac{r^4\epsilon^2\sin^2\phi}{a^2(1-\epsilon^2)^2}=-\frac{r^4}{a^2(1-\epsilon^2)}+\frac{2r^3}{a(1-\epsilon^2)}-r^2. \tag{31}$$

[3] The allowance to be made for the motion of the nucleus never differs perceptibly from that already made by introducing μ into equation (16), and the magnetic fields arising from the motions of the electron and of the nucleus are without perceptible effect (C. G. Darwin, *Phil. Mag. 39*, pp. 537-551; 1920). The correction which would be required if the nucleus or the electron were oddly shaped, if the nucleus were a magnet, or if there were entrainment of the potential energy of the system by the moving electron, have been evaluated by various people; consult A. E. Ruark, Astroph. Jl., 58, pp. 46-58 (1923).

[4] The ambiguity of sign which arises in the course of the development may be resolved by thinking of the limiting case of the circle ($\epsilon=O$).

All this is geometry. We must now prove that a particle moving under the influence of an inverse-square attraction, drawing it towards a fixed point, will describe an ellipse with that fixed point in one of its foci—will describe, otherwise expressed, a curve defined by equation (31).

As the particle is an electron, and the fixed point is occupied by a nucleus of charge E, the mutual attraction is eE/r^2 when their dis-

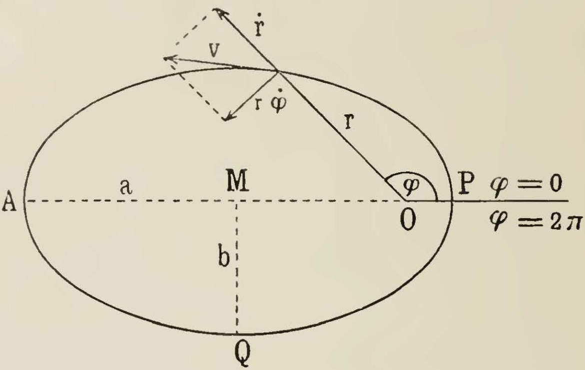


Fig. 2—Diagram to illustrate the notation used in describing elliptical orbits

tance apart is r. Equating this attraction to the product of the mass of the electron into the sum of its accelerations, linear and "centrifugal," we have

$$eE/r^2 = -m\frac{d^2r}{dt^2} + mr\left(\frac{d\phi}{dt}\right)^2 \tag{32}$$

It is necessary to assume the law of conservation of angular momentum; the angular momentum of the electron $mr^2\, d\phi/dt$ about the centre of attraction remains constant in time:

$$mr^2\frac{d\phi}{dt} = p, \tag{33}$$

inserting which into (32) we have

$$eE/r^2 = -m\frac{d^2r}{dt^2} + p^2/mr^3 \tag{34}$$

This is to be integrated in the usual way, by multiplying each term with $2(dr/dt)$; the result is

$$\left(\frac{dr}{dt}\right)^2 = -p^2/m^2r^2 + 2eE/mr - C, \qquad (35)$$

the last symbol standing for a constant of integration. Finally

$$\begin{aligned}(dr/d\phi)^2 &= (dr/dt)^2/(d\phi/dt)^2 = (dr/dt)^2(m^2r^4/p^2) \\ &= -Cmr^4/p^2 + 2eEmr^3/p^2 - r^2. \qquad (36)\end{aligned}$$

We recognize at once the identical form of this equation for the path in which the attracted particle moves and the equation (31) for the ellipse drawn about the centre of attraction as focus.

It remains only to identify the constants. Equating the coefficients of r^3 in the two equations, we have

$$p^2 = eEma\ (1-\epsilon^2). \qquad (37)$$

This is the equation giving the angular momentum of the electron in terms of the major axis and the eccentricity of the orbit. Equating the coefficients of r^4 in (31) and (36) we have

$$C = p^2/ma^2(1-\epsilon^2) = eE/a \qquad (38)$$

to determine the constant of integration in (35). If now the reader will take the expression for the energy of the system

$$W = \tfrac{1}{2}mv^2 - e^2/r = \tfrac{1}{2}m((dr/dt)^2 + r^2(d\phi/dt))^2 - e^2/r \qquad (39)$$

and substitute for $(d\phi/dt)$ according to (33) and for (dr/dt) according to (35) and (38), he should arrive at

$$W = -e^2/2a. \qquad (40)$$

This is the equation giving the energy of the system in terms of the constants of the ellipse; we see that the energy depends only on the major axis, not on the eccentricity, of the ellipse.

The period of revolution T is a little more difficult to calculate. The most logical procedure would be to take the reciprocal of the expression (35) for dr/dt, and integrate

$$t = \int(-p^2/m^2r^2 + 2eE/mr - eE/a)^{-1/2}dr \qquad (41)$$

around a complete revolution. The derivative dr/dt passes twice through zero in the course of the revolution, once at the point of the orbit nearest to the nucleus (perihelion) and once at the point farthest away. At these points $r = a(1\mp\epsilon)$, as can be seen from the geometry of the ellipse or by inserting these values into the expression for dr/dt.

By integrating (41) from one of these values to the other and doubling the result, we get the period of the revolution

$$T = 2\pi\sqrt{ma^3/eE}. \tag{42}$$

J2. Motion of an Electron in a Central Field Differing Slightly from an Inverse-square Field

Suppose we modify the atom-model composed of a nucleus and an electron by imagining that the force exerted by the one upon the other varies not exactly, but very nearly, as the inverse square of their distance apart. For instance, one might imagine that the force varies as $r^{2.001}$; or that the nucleus acts upon the electron with an attraction equal as heretofore to eE/r^2, *plus* an additional attraction (or repulsion) varying inversely as the cube of the distance. In any such case the potential energy of the atom-model would not be quite equal to $-eE/r$; there would be an additional term $f(r)$. In the case of an inverse-cube field superposed upon an inverse-square field, the expression for the potential energy would be

$$V = -eE/r - C/r^2 \tag{43}$$

The second term on the right hand side will be much smaller than the first, at and only at distances much greater than $2C/eE$; but by imagining C sufficiently small, we can arrange to have the inverse-cube field very much smaller than the inverse-square field, over all the region in which the orbit of the electron is likely to lie; and this is all that matters.

The orbit of the electron may be described, in all these cases in which the force deviates very slightly from an inverse-square force, as an *ellipse precessing in its own plane*. That is to say: an ellipse of which the major axis swings at a uniform rate around the nucleus as if around an axle perpendicular to its own plane—as though the electron were a car, running around and around an elliptical track, quite unaware that the track itself is endowed with a revolving motion of its own. (Or, in other and more sophisticated words, the orbit of the electron is an ellipse stationary in a coordinate-system revolving around the nucleus at a uniform rate). Such an orbit is known as a "rosette," and a part of a rosette is shown in Fig. 3.

Another way of describing the important feature of this orbit is to say that the two coordinates r and ϕ of the electron in its orbit (referred to O as origin and OP as the direction $\phi = 0$, in Fig. 3), while they are both periodic, do not have the same period. While r is running through its entire cycle from $r_{max.}$ to $r_{min.}$ and back again, the electron is moving from one point of tangency with the dashed circle, inward around the nucleus, back to the next point of tangency; meanwhile, ϕ is running through an entire circuit amounting to 2π, and in addition through the angle $\Delta\phi$. Thus the period T_r of r stands to the period T_ϕ of ϕ as

$$T_r : T_\phi = \frac{2\pi+\Delta\phi}{2\pi} = \frac{2\pi+2\pi\omega T_r}{2\pi} \tag{11}$$

in which expression the symbol ω stands for the frequency of the precession (i.e., the reciprocal of the time the major axis requires to

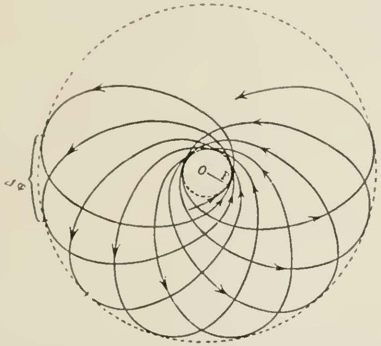


Fig. 3—Rosette orbit, resulting from a precession superposed upon an elliptical orbit

trace out the entire dashed circle). One might say that the two frequencies $\omega_r = 1/T_r$ and $\omega_\phi = 1/T_\phi$ are slightly out of tune with one another. So long as the force acting upon the electron is exactly an inverse-square force, these two frequencies are perfectly in tune, the ellipse is stationary; when the inverse-square force is slightly altered, the two frequencies fall out of tune and the ellipse revolves. In general, the two frequencies will be incommensurable with one another; the rosette will never return into itself, the electron will go on winding its path over and over and over the interior of the dashed circle, passing eventually within any assignable distance, no matter how small, of any point selected at random, and "covering the interior of the circle everywhere dense" as the mathematicians say. Therefore, although the variables r and ϕ are individually periodic, the motion of the electron never quite repeats itself. Such a system is called *conditionally periodic*.

When we come to consider the atom-models proposed for atoms with more than one electron, we shall make use of these ideas; but that will not occur before the Third Part of this article. However, one application can be made to the theory of hydrogen and ionized helium.

J3. *Motion, in an Inverse-square Central Field, of an Electron of Which the Mass Varies as Prescribed by the Theory of Relativity*

According to "relativistic mechanics," as distinguished from "Newtonian mechanics," the mass m of an electron (or anything else) varies with its speed v in the manner

$$m = m_o / \sqrt{1 - v^2/c^2} \tag{45}$$

and the force F acting upon it produces an acceleration dv/dt given not by the familiar equation *force = mass × acceleration*, but by the equation

$$F = d(mv)/dt \tag{46}$$

If we suppose the electron revolving in a perfect inverse-square field about the nucleus, and apply these equations of relativistic mechanics, we arrive at the same result as though we had used the equations of Newtonian mechanics, but had assumed that the field acting upon the electron is the sum of an inverse-square attraction and an inverse-cube attraction. Specifically, the result is formally equivalent to the result attained by continuing to use Newtonian mechanics, and assuming that the potential energy of the atom-model is given by (43) with the following value inserted for the constant C:

$$C = -e^2E^2/2m_oc^2 \tag{47}$$

The orbit is a rosette; and all the general remarks made in section J2 about rosette orbits may be repeated for this case.

J4. *Motion of an Electron in a Field Compounded of an Inverse-square Central Electric Field and an Uniform Magnetic Field*

Here we have a famous theorem of Larmor's to help us. According to this theorem, a magnetic field H acting upon a revolving electron, or a system of revolving electrons, produces no other effect than a precession of the entire system about the direction of the magnetic field at the frequency

$$\omega_L = eH/4\pi mc \qquad (18)$$

In other words, the motion of the electron or electrons is, when referred to a coordinate system revolving about the direction of the field with frequency $eH/4\pi mc$, the same as without the field it would be, when referred to a stationary coordinate system.

If the field happens to be normal to the plane of an elliptical orbit being described by an electron about a nucleus, the ellipse will be transformed into a rosette. If the field is neither exactly normal nor exactly parallel to the plane of the ellipse, this plane may be imagined to swing around the direction of the field (around the line through the nucleus parallel to the field) like a precessing top, carrying the orbit with it.

These statements are inexact if the rate of precession so calculated is not quite small in comparison with the rate of revolution of the electron.

15. *Motion of an Electron in a Field Compounded of an Inverse-square Central Electric Field and an Uniform Electric Field*

This problem may be regarded as the limiting case of a more general problem phrased as follows: to determine the motion of a particle attracted by two fixed points according to the inverse-square law. Imagine one of the fixed points to recede to infinity, its attracting-power meanwhile rising at the proper rate to keep the field in the region of the other at a finite value; and you have the case described in the sub-title above. Jacobi solved the general problem a century or so ago.

The motion is difficult to realize and impossible to describe in words, and seems also to be impossible to represent by any adequate two-dimensional sketch. The electron makes circuits around the line through the nucleus parallel to the uniform field, and in each circuit it describes a curve which is very nearly an ellipse; but the consecutve loops, as in the case of Fig. 3, do not coincide; furthermore, they are not alike in shape, and they are not plane. The electron winds around and around through the volume of what I am tempted to call a doughnut, surrounding the aforesaid line as its axis; and in the course of time its path fills up the doughnut "everywhere dense," as the path of the electron in Fig. 3 would fill up the interior of the dashed circle.

I hope it will be appreciated that the foregoing statements about the orbits are fatally incomplete, except in the first case. Nothing could be done unless it were possible to know, not merely the general shape of each type of orbit, but the exact mathematical expression for it, and for the energy-value of each orbit of each type. In some cases this knowledge is available; in others, it is not. For the cases designated here by J3, J4 and J5, it is available; wherefore it is possible to go about the process of seeking the distinctive features of orbits possessing the preassigned energy-values, or inversely the energy-values of orbits distinguished by certain features.

K. Further Interpretation of the Spectra of Hydrogen and Ionized Helium

Continuing for the moment to accept the energy-values of the stationary states of the hydrogen atom as given by

$$W_1 = -Rh,\ W_2 = -Rh/4,\ W_3 = -Rh/9,\ \ldots\ .$$

and continuing to accept the atom-model consisting of a nucleus and a revolving electron; let us consider what are the properties of the *elliptical* orbits, in which if the electron revolved, the atom-model would possess one or another of the required energy-values.

According to equation (40), the energy of the atom-model, when the electron is revolving in an ellipse of which the major axis is $2a$, is given by

$$W = -eE/2a$$

irrespective of the eccentricity of the ellipse. In this, as in all following equations, E is equal to e for hydrogen and to $2e$ for ionized helium. If we set this expression equal to one of the required energy-values, for instance to W_1, we have

$$2a_1 = -eE/W_1 = eE/Rh. \tag{50}$$

The atom-model therefore has the proper energy-value W_1 for the normal state of the hydrogen atom, if the electron is revolving in *any* ellipse for which the major axis is eE/Rh. The circle of diameter eE/Rh of which we have heretofore been thinking is only one of these ellipses, it is the one for which the major and the minor axes are identical and $\epsilon = 0$; there is an infinity of others.

Should we then divest the circular orbits of the prominence which has been accorded to them, and assume for instance that when the atom is in its normal state the electron is moving in any one of the infinity of ellipses of which the major axis is eE/Rh? This might be dangerous, for we have identified certain distinctive features of the permissible circular orbits which may be essential; and these features may not be transferable to the ellipses. Let us test them.

The second and the third of the three distinctive features which I cited are transferable that is they can be extended to the totality of all ellipses having one or another of the energy-values $-Rh/n^2$, and they differentiate these from all other ellipses. For it can be shown, by integrating the kinetic energy K (the first term on the right hand side of (39)) around an elliptical orbit, that

$$I = \int 2K\,dt = 2\pi\sqrt{ameE} \qquad (51)$$

depending only on the major axis a of the orbit. Now we have shown that $I = nh$ for the nth of the permissible circles; hence for each ellipse having the same major axis as the nth permissib.e circle, in other words for each ellipse of energy-value $-Rh/n^2$, we have

$$I = nh$$

and the second of the distinctive features is transferable to the ellipses. It is the same for the third; for T is by (42) dependent on a only, and so

$$Lim\ \omega = Lim\ \nu.$$

But it is otherwise with the first.

In the first place it was shown that the angular momentum of the electron in the circle of diameter eE/Rh is equal to $h/2\pi$. Obviously this cannot be true of all the ellipses of major axis eE/Rh. For according to (37), the angular momentum of the electron in such an ellipse is

$$p = \sqrt{eEma(1-\epsilon^2)} \qquad (52)$$

depending on the eccentricity. This is equal to $e\sqrt{ma}$, which by (12) is equal to $h/2\pi$, only if $\epsilon = 0$. The circle therefore is the only orbit for which the energy-value and the angular momentum of the atom are simultaneously equal to $-Rh$ and to $h/2\pi$ respectively. If we admit the ellipses to equal value with the circle, we concede that the equality of the angular momentum with $h/2\pi$ is of no significance.

There is a partial escape from this conclusion for the remaining stationary states. Take for instance the second, of energy-value $-Rh/4$. The circular orbit of diameter $4eE/Rh$, for which the atom possesses this energy-value, is distinguished by the angular momentum $2h/2\pi$. For each of the infinity of ellipses possessing the same major axis $4eE/Rh$ there is a different value of the angular momentum;

but there is one among them for which the angular momentum is equal to $h/2\pi$. And in general for the nth stationary state of energy-value $-Rh/n^2$, there are n elliptical (including one circular) orbits which would give the same energy-value and n values of angular

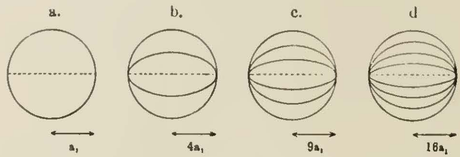


Fig. 4a—Diagram to show the proportional dimensions of ellipses with identical total quantum-number $n=I/h$ and different azimuthal quantum-numbers $k=1, 2 \ldots n$. From left to right we have the cases $n=1, 2, 3, 4$, on scales varying as indicated by the subjoined arrows.

momentum equal respectively to $nh/2\pi$, $(n-1)h/2\pi, \ldots h/2\pi$. These, as the reader can show from (52), are distinguished by the following values of ϵ:

$$\sqrt{1-\epsilon^2}=k/n \qquad k=1, 2 \ldots n. \tag{53}$$

Thus if we desire to regard the equality of angular momentum with an integer multiple of $h/2\pi$ as being essential to the permissible orbits, we can keep, along with the circles, some of the other elliptical orbits compatible with the prescribed energy-values; but except for these

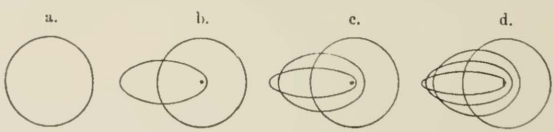


Fig. 4b—The same ellipses as appear in Fig. 4a, drawn confocally as they should appear, instead of concentrically

few, the infinity of elliptical orbits will remain unavailable. There is additional reason for liking to do this; for it amounts to a quite natural generalization of the condition imposed on the angular momentum, which as we saw it is highly desirable to generalize if possible. The angular momentum $mr^2(d\phi/dt)$, which I shall hereafter call $p\phi$ instead of simply p, stands on an equal footing with the radial momentum $pr=m(dr/dt)$ of the electorn; in the Hamiltonian equations for the motion of the particle, these two quantities stand side by side. Now the condition imposed upon the angular momentum

$p\phi$ of the electron in its various circular orbits is $p\phi = nh/2\pi$, which may be written

$$\int_0^{2\pi} p_\phi \, d\phi = nh \tag{54}$$

the integral being taken around a complete revolution, a formulation in which the somewhat distressing factor $1/2\pi$ conveniently vanishes. Corresponding to this integral we have another

$$\int p_\phi \, dr = m\int \frac{dr}{d\phi} d\phi \tag{55}$$

also to be taken around a complete revolution, therefore from $r_{min.} = a(1-\epsilon)$ to $r_{max.} = a(1+\epsilon)$ and back again. The materials for performing this integration are furnished in equation (35); if the reader can perform it he will arrive at the value.

$$\int p_r \, dr = 2\pi p_\phi \left[\frac{1}{\sqrt{1-\epsilon^2}} - 1 \right] \tag{56}$$

and if the eccentricity of the ellipse conforms to equation (53), so that the integral of the angular momentum of the electron is kh, then the integral of the radial momentum is

$$\int M_r dr = (n-k)h. \tag{57}$$

Our position may now be described in the following words. We have accepted the values $-Rh/n^2$ ($n = 1,2,3 \ldots$) for the successive stationary states of the hydrogen atom; we have accepted an atom-model consisting of a nucleus and a revolving electron; we have traced the orbits which would entail these various energy-values, and we have found that for each of these energy-values there are infinitely many elliptical orbits which would entail it,—to wit, for the nth stationary state, all the infinitely many ellipses of which the major axis is given by

$$2a_n = n^2h^2/2\pi^2 meE. \tag{58}$$

Furthermore we have sought for distinctive features which might discriminate these ellipses from all the others which entail "wrong" energy-values, i.e., energy-values which are not included in the list $-Rh$, $-Rh/4$, $-Rh/9$ .... One such we found in the integral $\int 2K dt$ of the kinetic energy of the electron around the ellipse; this integral assumes the value nh for each ellipse which entails the energy-value $-Rh/n^2$, so that we could define the permitted orbits as those for which $\int 2Kdt$ = any integer multiple of h. Another such distinctive feature we found in what was expressed by the equation (23) $Lim\ \omega = Lim\ \nu$. First of all, however, we tried to apply a principle of the effect that the angular momentum of the atom when the electron is revolving in one of the permitted orbits must be an integer multiple of $h/2\pi$. We found, in essence, that this attempt amounted to picking out for each of the prescribed energy-values, one or several out of the infinity of elliptical orbits which would entail it, and eliminating all the rest. But is there sufficient reason for doing a thing like this?

Apparently there is; and the reason for so believing lies precisely in the details of the hydrogen spectrum which I have hitherto passed over—in the doubleness of the lines of the Balmer series, which shows that instead of a stationary state of energy-value $-Rh/4$ there are two stationary states of which the energy-values lie extremely close to one another and to this value, and which suggests that the other stationary states may likewise be resolvable into groups of stationary states (a suggestion confirmed by the spectrum of ionized helium). At the beginning, let us consider only the state of which the energy-value is $-Rh/4$. We have seen that this is the energy-value corresponding to any and every one of the elliptical orbits of which the major axis is

$$2a_2 = 4h^2/2\pi^2\ meE \tag{59}$$

among which infinity of elliptical orbits, there is just one (a circle) for which the angular momentum of the atom is $2h/2\pi$, and just one other for which it is $h/2\pi$, and no others for which it is any integer multiple of $h/2\pi$ at all. But these two, like all the rest characterized by (58), entail the same energy-value and so are indistinguishable among the crowd if every one of our assumptions is absolutely true. But if one of them should deviate slightly from the truth—if for instance the law of force between the nucleus and the electron should deviate slightly from the inverse-square law, or if a small extraneous force should be impressed upon the atom, or if the mass of the electron should slightly vary as it revolves in its orbit—then we have seen that all the orbits would be altered, and these two orbits may be so altered as to be distinguishable from the rest. And this in fact is what appears to be responsible for the fine structure of the hydrogen and ionized-helium. Owing to the variation of the mass of the electron, with its speed, each ellipse is transformed into a rosette; and though the energy-values of all the ellipses would be equal, the energy-values of the rosettes are not.

Let us now reverse the procedure of the foregoing paragraphs. Instead of asking what is the angular momentum of the atom when the electron is revolving in such an orbit that the energy of the atom is $-Rh/4$, let us ask what is the energy of the atom when the electron is revolving in a rosette such that the angular momentum of the atom is $2h/2\pi$. It is best to put the question thus: what is the energy of the atom when the electron is revolving in a rosette[10] such that the integral of the angular momentum around a revolution is $2h$?

$$\oint p_\phi \, d\phi = 2h. \tag{61}$$

The energy-value in question, which I designate by W_{22} for a reason which will presently appear, is found by calculation to be

$$W_{22} = -Rh/4 - Rh\alpha^2/64 \tag{62}$$

in which α is a symbol meaning

$$\alpha = 2\pi e^2/hc = 7.29 \cdot 10^{-3}. \tag{63}$$

(This expression incidentally is not the exact consequence of the equations of the motion, but an approximation to it, quite sufficiently accurate under these circumstances). Next let us ask what is the energy of the atom when the electron is revolving in a rosette[10] such that

$$\oint p_\phi \, d\phi = h. \tag{64}$$

Calling this energy-value W_{21}, it is calculated that

$$W_{21} = -Rh/4 - Rh5\alpha^2/64. \tag{65}$$

Incidentally it is found, as in the previous simpler case, that when $\oint p_\phi d\phi = h$, then also $\oint p_r dr = h$.

The energy-values corresponding to the two orbits defined by (68) and (71) therefore differ by the very small amount

$$W_{22} - W_{21} = -Rh\alpha^2/16 = -Rh(3.32 \cdot 10^{-6}). \tag{66}$$

I said at first that the various "lines" of the Balmer series in the spectrum of hydrogen correspond to transitions into the stationary state of energy-value $-Rh/4$ from other stationary states; and that unusually good spectroscopes show each of these lines to be a pair of lines very close together. May this be explained by the theory culminating in equation (66)? If so, the frequency-difference between the two lines of each doublet must be the same, and equal to

[10] This rosette is degenerated into a circle; the precession amounts effectively to an additional term in the expression for the angular velocity of the electron.

$(W_{22}-W_{21})h = R\alpha^2/16 = 1.09 \cdot 10$. The wave-length difference, which is the quantity directly measured by spectroscopists, varies from one doublet to another; for the first doublet of the Balmer series, known as $H\alpha$, the mean wavelength of which is $6.563 \cdot 10^{-5}$ cm., it should be equal to $1.58 \cdot 10^{-9}$ cm.

Many independent measurements of these wavelength differences have been made, most of them upon the first doublet of the series, a few upon other doublets as far along as the fifth. Some were made long before, others after Sommerfeld published the foregoing theory. The various values found for the various wavelength-differences have all been within 20% of the value required by equation (66); within this range they have fluctuated, one or two spectroscopists of repute have maintained that the actual values are unmistakably different from the computed value; but the balancing of evidence now seems to point more and more closely to the desired value as the right one [11].

This prediction of the wavelength-differences between the components of the doublets which make up the Balmer series may be taken tentatively as the third of the numerical agreements which fortify Bohr's atom-model. So taking it, let us generalize the theory to the full extent already suggested. Returning for a moment (merely for ease of explanation) to the over-simplified case of an atom consisting of a nucleus and a revolving electron of which the mass does not vary with its speed: we saw that the energy-value $-Rh/n^2$ is entailed by each and every one of the n elliptical orbits for which the integral of the angular momentum and the integral of the radial momentum are given by assigning the n values $k=1, 2, 3 \ldots n$ to the symbol k in the following equations:

$$\int p_\phi d\phi = kh, \int p_r dr = (n-k)h. \qquad (67)$$

This I will express in another way by saying that the energy-value $-Rh/n^2$ is entailed by each of the n orbits having the *azimuthal*

[11] This is one of those embarrassing questions as to which the experimental doctors still disagree, making it folly indeed for anyone else to pretend to decide. The three latest measurements, which are those of Shrum, Oldenberg, and Geddes, agree passably with the value resulting from the theory I have presented. Yet Gehrcke and Lau defend their measurements, made in 1920 and 1922, which give values about 20% too low; and Gehrcke at least is an authority to whom lack of experience in this field certainly cannot be imputed. I evade this issue by referring the reader to the articles by Shrum (*Proc. Roy. Soc.* A105, pp. 259-270; 1923) for the bibliography of earlier work and the account of the latest; of Ruark (*l. c. supra*) for the contention that the data sustain the theory; of Lau (*Phys. ZS.* 25, pp. 60-68; 1924) for the contrary contention.

The issue is further complicated by the predictions quoted in the next paragraph above, although not seriously enough to disqualify the foregoing remarks.

quantum-numbers $k=1, 2, \ldots n$; meaning by azimuthal quantum-number the quotient of $\int p_\phi d\phi$ by h. If now we take account of the variation of the mass of the electron with its speed, and calculate the energy-values for the n rosettes obtained by assigning the values $1, 2, 3 \ldots n$ successively to the symbol k in (67), we shall find that these n energy-values are all distinct, deviating slightly from $-Rh/n^2$ and from each other. Therefore, there should be three stationary states of energy-values W'_{33}, W'_{32}, W'_{31}, all differing by a little from $-Rh/9$ and from each other; there should be four stationary states of energy-values W'_{44}, W'_{43}, W'_{42}, W'_{41}, all nearly but not quite equal to $-Rh/16$ and each other; and so forth. (The reason for such symbols as W'_{21} will now appear; the first subscript represents the total, the second the azimuthal quantum-number of the orbit in question.) In general there are n stationary states in the group corresponding nearly to the mean energy-value $-Rh/n^2$; and the expressions for their several values are obtained by putting k equal to the various values $1, 2, 3 \ldots n$ in the formula.

$$E=-Rh/n^2\left[1+\frac{\alpha^2}{n^2}\left(\frac{n}{k}-\frac{3}{4}\right)\right]. \tag{68}$$

Owing to these complexities the lines of the Balmer series should be not doublets, but groups of many more lines; e.g., the transitions from what I had called the stationary state of energy-value $-Rh/9$ to the stationary state of energy-value $-Rh/4$ are transitions of six sorts, from each of three initial states to each of two final; and the first "line" of the Balmer series might be expected to be sextuple.

The trial of these ideas is best made upon the spectrum of ionized helium. The separation between the energy-values of stationary states sharing the same total quantum-number and differing in azimuthal quantum-number is increased, when we pass from an atom-model in which the charge on the nucleus is e to one in which it is Ze, in the ratio Z^4:1; in this instance 16:1. The system of component lines, or the so-called "fine structure" to be expected for any "line" of the hydrogen spectrum should be spread out on a scale sixteenfold as great for the corresponding "line" of the ionized-helium spectrum. The trial was made by Paschen; the comparison between the fine structure of several of the "lines" of ionized helium and the components to be expected from the foregoing theory, yielded what appear to be very satisfactory results. This matter I discussed over several pages of the First Part of this article; and for economy of space I refer the reader back to them, and at this place say only that the "other numerical agreements between the production and the data" to which I there allude, are agreements of the same character as the agreement between the spacing of the component lines of the Balmer series doublets, and the numerical value of the expression in equation (73). That is to say: the pattern of the fine structure, into which by a good spectroscope the lines of ionized helium are resolved, agrees more or less with the pattern to be expected from the theory, not only in appearance but in scale. Combining these agreements with the other one, we are probably justified in counting the latter as the third of the conspicuous numerical agreements which make Bohr's atom-model plausible [12].

Now let us examine the situation again. (Considering the abstruseness of these matters, I hope that few readers will resent these frequent repetitions of past remarks.) Accepting for the atom of hydrogen (and of ionized helium) an atom-model consisting of a nucleus and an electron, we have traced orbits for the electron such as entail energy-values for the atom equal to those of the known stationary states. At first we ignored both the experimental fact that the lines of hydrogen and those of ionized helium have a fine structure, and the theoretical likelihood that the mass of the electron varies with its speed; and we found that the orbits are ellipses. Later on, we took cognizance of both these things; and we found that the orbits are rosettes. Yet merely to trace the orbits which yield the required energy-values, the so-called "permissible" orbits, amounts to little. It is essential to find distinctive features which set the permissible orbits apart from all the others—on success in achieving this, the whole value of the theory depends.

Now at the very beginning it was shown that, if we ignore the variation of the mass of the electron with its speed, and if we consider circular orbits only—then the permissible circular orbits which yield the required energy-values $-Rh/n^2$ of the stationary states (fine-structure being ignored!) are those for which

$$\oint p_\phi d\phi = nh \tag{69}$$

in which equation p_ϕ stands for the angular momentum of the motion, and n for any positive integer; and the integral is taken around a complete cycle of ϕ.

[12] For the experimental results and the comparison of data with predictions see Paschen's great paper (*Ann. d. Phys.* 50, pp. 901-940; 1915) which however is anything but easy to read, so that Sommerfeld's presentation will probably be preferred; likewise Birge's article (*Phys. Rev.* 17, pp. 589 ff, 1921) to which the same words apply. The agreements are impressive. On the other hand I note that Lau (*l. c. supra*) concludes from the same data that there is a disagreement between data and predictions, in the same sense and of about the same magnitude as the disagreement which he claims to occur in the hydrogen spectrum.

It was next shown that when we make allowance for the variation of the mass of the electron with its speed, then the permissible rosette orbits which yield the required energy-values of the stationary states (fine structure being taken into account[1]) are those for which

$$\int p_r dr = n_1 h \qquad \int p_\phi d\phi = n_2 h \tag{70}$$

in which equations p_r and p_ϕ stand for the radial and angular momenta—the momenta belonging to the variables r and ϕ respectively—and n_1 and n_2 for any positive integers; and the integrals are taken around complete cycles of r and ϕ respectively.

The equations (70) look like a very natural and pleasing generalization of the equation (69). It is possible to go somewhat further. Consider that, when the electron was supposed to move in a circle, its position was defined by one variable ϕ; and the permissible circles were determined by one integral. Further, when the electron was supposed to move in a rosette, its position was defined by two variables r and ϕ; and the permissible rosettes were determined by two integrals. Now when the electron is subjected, for instance, to an uniform magnetic field superposed upon the field of the nucleus, its motion is three-dimensional. Three variables are required to define its position; for instance, the variables r, θ and ψ of a polar coordinate system with its polar axis parallel to the direction of the magnetic field. Three corresponding momenta p_r, p_θ and p_ψ can be defined. It seems natural to generalize from (69) through (70) to a triad of equations, and say that the permissible orbits are those for which

$$\int p_r dr = n_1 h \qquad \int p_\theta d\theta = n_2 h, \qquad \int p_\psi d\psi = n_3 h \tag{71}$$

in which equations n_1, n_2, n_3 all stand for positive integers, and the integrals are taken around complete cycles of r, θ and ψ respectively.

When this is done for the specific case of an electron moving under the combined influence of a uniform magnetic field and the field of a nucleus, the result is entirely satisfactory. That is to say: when the permissible orbits are determined by using the equations (71) upon the general type of orbit described in section J I, and when their energy-values are calculated, it is found that they agree very well with the observed energy-values of the stationary states of hydrogen in a magnetic field. This may be regarded as the fourth of the numerical agreements which fortify Bohr's atom-model. As I shall end this part of the present article by a presentation of the effect of the magnetic field made in a somewhat different manner, I reserve the details for the following section.

Yet it cannot be said that equation (71) is the utterance of the much-desired General Principle, of the distinctive feature *par excellence* which sets all permissible orbits apart from all non-permissible orbits in every case. The most that can be said is this, that equation (71), if properly interpreted, is the widest partial principle that has yet been discovered. But it suffers limitations. I do not mean, as might be thought, that cases have been discovered in which the permissible orbits determined by such equations as (71) have energy-values not agreeing with those of the observed stationary states. The difficulty is, that equations such as (71) cannot even be formulated in many cases, because the necessary mechanical conditions do not exist.

This matter is a hard one to make clear; but the limitation can be at least partially expressed in the following way. Revert to the equations (70) which were applied to the rosette orbits. The first of the integrals in (70) is to be taken over an entire cycle of the variable r. Now it was said in section J2 that the periods of the two variables r and ϕ are not equal, and in general they are incommensurable. When the variable r describes a complete cycle, r and dr/dt both return to their initial values; but ϕ and $d\phi/dt$ do not have, at the end of the cycle of r, the same values as they had at its beginning. It follows that if p_r depends on ϕ or on $d\phi/dt$, the first of the two integrals in equation (70) will have different values for different cycles of r. If so, the conditions imposed upon the permissible orbits by (70) would have no meaning. The conditions have a meaning, only if each of the integrals in (70) has the same value for every cycle of its variable—therefore, only if p_r depends on r only, and p_ϕ depends on ϕ only. And in general, such a set of equations as (71) has a meaning, only if it is possible to find a set of variables such that the momentum corresponding to each of them depends on and only on the variable to which it corresponds; or, in technical language, only if it is possible to effect *separation of variables.*

Separation of variables is possible in some cases, and in others it is not. When the periods of all the variables are equal, as they are when we imagine an electron of changeless mass revolving in an inverse-square field, it is clearly always possible; the difficulty described in the foregoing paragraph does not occur. In the other cases which I have outlined—when the electron is imagined to move in an inverse-square field according to the laws of relativistic mechanics, and when it is imagined to move in a field compounded of an inverse-square field and an uniform magnetic field separation of variables is possible. For these cases, therefore, the conditions (70) and (71) are applicable, and have meaning.

There is one other important case in which it is possible so to select the variables that separation can be effected. This is the case of an electron moving according to the laws of Newtonian mechanics in a field compounded of an inverse-square field and an uniform electric field. Although the motion is three-dimensional, and three coordinates are required and suffice to determine it, these three coordinates may not be chosen at random; and the three obvious ones would be worthless for our purpose. If we should choose the polar coordinates r, θ, and ψ employed in formulating the equations (71), we should find that the momenta p_r, p_θ and p_ψ do not depend each exclusively upon the variable to which it corresponds. The procedure to be followed is anything but obvious; but Jacobi found that if paraboloidal coordinates are used instead of polar, separation of variables can be effected. One must visualize two families of coaxial and confocal paraboloids, their common focus at the nucleus, their noses pointing in opposite directions along their common axis which is the line drawn through the nucleus parallel to the electric field. The position of any point through which the electron may pass is given by the parameters ξ and η of the two paraboloids which intersect at that point, and by an angle ϕ defining its azimuth in the plane normal to the axis, quite like the angle ψ of a system of polar coordinates. When the motion of the electron is expressed in terms of these coordinates, the corresponding momenta p_ξ and p_η depend only upon ξ and η respectively and p_ϕ is constant; hence the integrals taken over cycles of ξ, η, and ϕ respectively, on the right-hand sides of the equations,

$$\int p_\xi d\xi = n_1 h, \qquad \int p_\eta d\eta = n_2 h \qquad \int p_\phi d\phi = n_3 h \tag{72}$$

have definite meanings, and the equations themselves define particular orbits. Epstein determined the orbits defined by these equations, and calculated their energy-values. These agreed well with the energy-values of the stationary states of hydrogen in an electric field, inferred from its spectrum. This is the fifth of the striking numerical agreements upon which the credit of Bohr's atom-model chiefly depends [13].

[13] See Epstein's article (*Ann. d. Phys. 50*, pp. 489-520; 1916), or the more perspicuous account by Sommerfeld, in which it is stated that the pattern of the components into which the first four lines of the Balmer series are resolved by the electric field agrees with the predictions so far as the number and relative spacings of the components are concerned; while to attain agreement in regard to the absolute spacings, it is necessary only to assume that Stark's estimate of the field was 3% in error, which is quite easy to accept.

It is important to note that if we had made allowance for the variation of the mass of the electron with its speed—if in other words we had used the equations of relativistic mechanics, which are probably the right ones to use—separation of variables could not have been effected either in this paraboloidal coordinate-system, or in any other. Yet the stationary states are found by experiment to be sharply defined, and to have approcimately the energy-values determined by (72). This can mean only that the desired General Principle for determining the permissible orbits is not completely expressed by such sets of equations as (71) or (72). Those equations are valid only for systems of a certain kind (those for which separation of variables is possible). The General Principle must be valid for systems of this kind and the other kind as well. For systems of this kind, it must become equivalent with the conditions formulated in (71) and (72)—the General Quantum Conditions for Separable Systems. Or at least, the results to which it leads must be indistinguishable from the results to which these lead. The General Principle for systems of every kind has not been discovered; perhaps it does not exist. Bohr is striving to infer it by generalizing from the third of the properties of the permissible circular orbits, which I mentioned in Section H and expressed by equation (23). He has attained some notable successes, which I hope that it will be possible to expound in the Third Part of the article.

L. Magnetic Properties of the Atom Model

After this rather arduous pilgrimage through a succession of abstract reasonings, the reader may welcome an account in simpler fashion of the manner in which Bohr's atom-model is adapted to explain the behavior of the atom in a magnetic field. This is an alternative method of arriving at the same results as are attained by means of equations (71).

It was stated in section E9 of the First Part of the article, that the spectrum of a radiating substance in a magnetic field indicates that the field acts by replacing each of the stationary states, which the substance possesses when there is no magnetic field prevailing, by two or more new stationary states. The energy of each of the new stationary states differs from that of the stationary state which it replaces, by the amount

$$\Delta U = seHh/4\pi mc \tag{73}$$

in which H stands for the magnetic field strength and s for an integer, which must possess two or more values spaced at intervals of one unit[14].

The atom-model which we have been discussing at such length consists of an electron circulating in an elliptical orbit about a stationary nucleus; the minor variations due to the variation of the mass m of the electron with its speed, and to the motion of the nucleus, are now of comparatively little importance. An electron circulating in a closed orbit with frequency ν passes ν times per second through any point of its orbit, so that the charge passing per second through any such point is equal to that which would pass, if a continuous current $I = e\nu/c$ (measured in electromagnetic units) were flowing around the orbit. Now a current I flowing continuously around the curve bounding an area A is equivalent—so far as its field at a distance goes—to a magnet, of which the magnetic moment M is directed normally to the plane of the curve and is equal in magnitude to IA. The area of an ellipse of which the major axis is denoted by a and the minor axis $b = a\sqrt{1-\epsilon^2}$ is equal to $\pi ab = \pi a^2\sqrt{1-\epsilon^2}$. Hence the magnetic moment of the atom-model is equal to

$$M = e\nu\pi a^2\sqrt{1-\epsilon^2}/c \tag{74}$$

Further we have seen, by equations (37) and (42), that the angular momentum of the electron in its orbit is equal to

$$p = 2\pi m\nu a^2\sqrt{1-\epsilon^2} \tag{75}$$

Consequently

$$M/p = e/2mc \tag{76}$$

a rather surprisingly simple relation!

Now when a magnet of moment M is placed in a magnetic field of field-strength H, it acquires a certain potential energy ΔU in addition to the intrinsic energy which it possesses when oriented normally to the field—which depends on the angle θ between the

[14] Unlike some of the preceding derivations, this theory is not essentially limited to the case of an atom-model consisting of a nucleus and one electron. If there are several electrons describing closed orbits, the Larmor precession affects them identically; or, otherwse put, the magnetic field treats the atom as a unit having an angular momentum and a magnetic moment equal respectively to the vectorial sums of the angular momenta and the magnetic moments of the individual electrons. In fact the best verification of (73) is obtained from the lines belonging to the singlet systems of certain metals, which display "normal" Zeeman effect—the effect to which this theory is adapted. With anomalous Zeeman effect, against which this theory is powerless, we are not now concerned. In the case of hydrogen, the effect is complicated by the fine structure of the lines. With small magnetic fields it is normal, at least so far as the observations go. Each of the two stationary states of which the energy-values are given by (62) and (65) is replaced by two or more, conforming to (73).

direction of its magnetic moment and the direction of the field, and is given by

$$\Delta U = MH \cos\theta \qquad (77)$$

According to equation (73), the observed stationary states of hydrogen atoms in a magnetic field have specific discrete energy-values. These must correspond to specific discrete values of the angle θ; *the orientation of the atom in the magnetic field must be constrained to certain particular directions*, an extraordinary idea! We ascertain these "permissible directions" by equating the two values of ΔU figuring in (73) and (77), obtaining

$$seh/4\pi mc = M \cos\theta \qquad (78)$$

into which we then insert the expression for M in terms of p:

$$sh/2\pi c = p \cos\theta \qquad (79)$$

We have experimented at length with the notion that the angular momentum p of the electron in its orbit is constrained to assume only such values as are integer multiples of $h/2\pi$; let it be introduced here also. If $p = kh/2\pi$, then

$$s = k \cos\theta \qquad (80)$$

The angle θ may assume only such values, as will give to the quantity $s = k \cos\theta$ two or more values, differing by one unit. For instance, if $k = 1$, the values $\theta = 60°$ and $120°$ will suffice.

This, the most spectacular of all the remarkable consequences of Bohr's interpretation of the stationary states, is also the only one which has ever been directly verified.

The verification has not been made upon hydrogen nor upon ionized helium, but upon the atoms of certain metals [15]. I shall therefore reserve the account of it for the following sections of the article, where also there are certain other reasons for desiring to put it. Nevertheless, the reader should be aware of it at this point.

[15] I gave an account of the earliest of these experiments in the first article of this series (This Journal, *2*, October, 1923; pp. 112-114). The subsequent experiments have added nothing fundamentally new.

(To be continued)

Electric Circuit Theory and the Operational Calculus

By JOHN R. CARSON

NOTE. This is the first of three installments by Mr. Carson which will embody material given by him in a course of lectures at the Moore School of Electrical Engineering, University of Pennsylvania, May, 1925. No effort has been spared by the author to make his treatment clear and as simple as the subject matter will permit. The method of presentation is distinctively pedagogic. To electrical engineers and to engineering instructors, this exposition of the fundamentals of electric circuit theory and the operational calculus should be of great value.—EDITOR.

FOREWORD

THE following pages embody, substantially as delivered, a course of fifteen lectures given during the Spring of 1925 at the Moore School of Electrical Engineering of the University of Pennsylvania.

After a brief introduction to the subject of electric circuit theory, the first chapters are devoted to a systematic and fairly complete exposition and critique of the Heaviside Operational Calculus, a remarkably direct and powerful method for the solution of the differential equations of electric circuit theory.

The name of Oliver Heaviside is known to engineers the world over: his operational calculus, however, is known to, and employed by, only a relatively few specialists, and this notwithstanding its remarkable properties and wide applicability not only to electric circuit theory but also to the differential equations of mathematical physics. In the writer's opinion this neglect is due less to the intrinsic difficulties of the subject than to unfortunate obscurities in Heaviside's own exposition. In the present work the *operational calculus* is made to depend on an integral equation from which the Heaviside Rules and Formulas are simply but rigorously deducible. It is the hope of the writer that this mode of approach and exposition will be of service in securing a wider use of the operational calculus by engineers and physicists, and a fuller and more just appreciation of its unique advantages.

The second part of the present work deals with advanced problems of electric circuit theory, and in particular with the theory of the propagation of current and voltage in electrical transmission systems. It is hoped that this part will be of interest to electrical engineers generally because, while only a few of the results are original with the present work, most of the transmission theory dealt with is to be found only in scattered memoirs, and there accompanied by formidable mathematical difficulties.

While the method of solution employed in the second part is largely that of the operational calculus, I have not hesitated to employ developments and extensions not to be found in Heaviside. For example, the formulation of the problem as a Poisson integral equation is an original development which has proved quite useful in the actual numerical solution of complicated problems. The same may be said of the Chapter on Variable Electric Circuit Theory.

In view of its two-fold aspect this work may therefore be regarded either as an exposition and development of the operational calculus with applications to electric circuit theory, or as a contribution to advanced electric circuit theory, depending on whether the reader's viewpoint is that of the mathematician or the engineer.

I have not attempted in the text to give adequate reference to the literature of the subject, now fairly extensive. In an appendix, however, there is furnished a list of original papers and memoirs, for which, however, no claim to completeness is made.

CHAPTER I

The Fundamentals of Electric Circuit Theory

While a knowledge, on the reader's part, of the elements of electric circuit theory will be assumed, it seems well to start with a brief review of the fundamental physical principles of circuit theory, the mode of formulating the equations, and some general theorems which will prove useful subsequently.

First, the *circuit elements* are resistances, inductances, and condensers. The network is a *connected* system of circuits or branches each of which may include resistance, inductance and capacitance elements together with mutual inductance, and mutual branches.

The equations of circuit theory may be established in a number of different ways. For example, they may be based on Maxwell's dynamical theory. In accordance with this method, the network forms a dynamic system in which the currents play the role of velocities. If we therefore set up the expressions for the kinetic energy, potential energy and dissipation, the network equations are deducible from general dynamic equations.

The simplest, and for our purposes, a quite satisfactory basis for the equations of circuit theory are found in Kirchhoff's Laws. These laws state that

1. The total impressed force taken around any closed loop or circuit in the network is equal to the potential drop due to (a) resistance, (b) inductive reaction and (c) capacitive reactance.

2. The sum of the currents entering any branch point in the network is always zero.

Let us now apply these laws to an elementary circuit in order to deduce the physical significance of the circuit elements.

Consider an elementary circuit consisting of a resistance element R, an inductance element L and a capacity element C in series, and let an electromotive force E be applied to this circuit. If I denote the current in the circuit, the resistance drop is RI, the inductance drop is LdI/dt and the drop across the condenser is Q/C where Q is the charge on the condenser. It is evident that Q and I are related by the equation $I=dQ/dt$ or $Q=\int I dt$. Now apply Kirchhoff's law relating to the drop around the circuit: it gives the equation

$$RI+LdI/dt+Q/C=E.$$

Multiply both sides by I: we get

$$RI^2+\frac{d}{dt}\frac{1}{2}LI^2+\frac{d}{dt}\frac{Q^2}{2C}=EI.$$

The right hand side is clearly the rate at which the impressed force is delivering energy to the circuit, while the left hand side is the rate at which energy is being absorbed by the circuit. The first term RI^2 is the rate at which electrical energy is being converted into heat. Hence the resistance element may be defined as a device for converting electrical energy into heat. The second term $\frac{d}{dt}\frac{1}{2}LI^2$ is the rate of increase of the magnetic energy. Hence the inductance element is a device for storing energy in the magnetic field. The third term $\frac{d}{dt}Q^2/2C$ is the rate of increase of the electric energy. Hence the condenser is a device for storing energy in the electric field.

In the foregoing we have isolated and idealized the circuit elements. Actually, of course, every circuit element dissipates some energy in the form of heat and stores some energy in the magnetic field and some in the electric field. The analysis of the actual circuit element, however, into three ideal components is quite convenient and useful, and should lead to no misconception if properly interpreted.

Now consider the general form of network possessing n independent meshes or circuits. Let us number these from 1 to n, and let the corresponding mesh currents be denoted by $I_1, I_2 \ldots . I_n$. Let electromotive forces $E_1, E_2 \ldots . E_n$ be applied to the n meshes or circuits respectively. Let L_{jj}, R_{jj}, C_{jj} denote the total inductance, resistance and capacity in series in mesh j and let L_{jk}, R_{jk}, C_{jk} denote the corresponding mutual elements between circuit j and k. Now write down Kirchhoff's equation for any circuit or mesh, say mesh 1; it is

$$\left(L_{11}\frac{d}{dt}+R_{11}+\frac{1}{C_{11}}\int dt\right)I_1+\left(L_{12}\frac{d}{dt}+R_{12}+\frac{1}{C_{12}}\int dt\right)I_2+ \\ \dots+\left(L_{1n}\frac{d}{dt}+R_{1n}+\frac{1}{C_{1n}}\int dt\right)I_n=E_1$$

Corresponding equations hold for each and every one of the n meshes of the network. Writing them all down, we have the system of equations

$$\begin{array}{l} \left(L_{11}\frac{d}{dt}+R_{11}+\frac{1}{C_{11}}\int dt\right)I_1+\dots+\left(L_{1n}\frac{d}{dt}+R_{1n}+\frac{1}{C_{1n}}\int dt\right)I_n=E_1 \\ \text{-------------------------------------} \\ \left(L_{n1}\frac{d}{dt}+R_{n1}+\frac{1}{C_{n1}}\int dt\right)I_1+\dots+\left(L_{nn}\frac{d}{dt}+R_{nn}+\frac{1}{C_{nn}}\int dt\right)I_n=E_n \end{array} \quad (1)$$

The system of simultaneous differential equations (1) constitute the canonical equations of electric circuit theory. The interpretation and solution of these equations constitute the subject of Electric Circuit Theory, and it is in connection with their solution that we find the most direct and logical introduction to the Operational Calculus.

As an example of the appropriate mode of setting up the circuit equations, consider the two mesh network shown in sketch 1. Writing down Kirschhoff's Law for meshes 1 and 2, respectively, we have

$$\left(L_1\frac{d}{dt}+R_1+\frac{1}{C_1}\int dt\right)I_1+M\frac{d}{dt}I_2=E_1$$

$$+M\frac{d}{dt}I_1+\left(L_2\frac{d}{dt}+R_2+\frac{1}{C_2}\int dt\right)I_2=E_2$$

In this case the self and mutual coefficients are given by

$L_{11}=L_1$	$L_{22}=L_2$	$L_{12}=L_{21}=+M$
$C_{11}=C_1$	$C_{22}=C_2$	$C_{12}=C_{21}=o$
$R_{11}=R_1$	$R_{22}=R_2$	$R_{12}=R_{21}=o$

The conventions adopted for the positive directions of currents and voltages are indicated by the arrows. The sign of the mutual inductance M will depend on the relative mode of winding of the two coils.

Now write down Kirchhoff's Law, or the circuital equation for the network of sketch 2. They are

$$\left\{ (L_1+L_3)\frac{d}{dt}+(R_1+R_3)+\left(\frac{1}{C_1}+\frac{1}{C_3}\right)\int dt \right\} I_1$$

$$-\left(L_3\frac{d}{dt}+R_3+\frac{1}{C_3}\int dt\right) I_2 = E_1,$$

$$-\left(L_3\frac{d}{dt}+R_3+\frac{1}{C_3}\int dt\right) I_1$$

$$+\left\{ (L_2+L_3)\frac{d}{dt}+(R_2+R_3)+\left(\frac{1}{C_2}+\frac{1}{C_3}\right)\int dt \right\} I_2 = E_2.$$

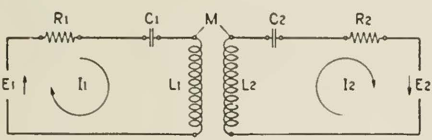
Comparison with equations (1) shows that

$$L_{11}=L_1+L_3 \qquad L_{22}=L_2+L_3 \qquad L_{12}=L_{21}=-L_3$$

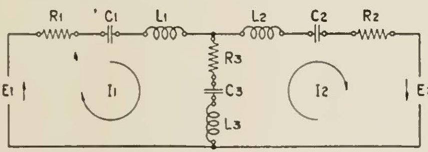
$$R_{11}=R_1+R_3 \qquad R_{22}=R_2+R_3 \qquad R_{12}=R_{21}=-R_3$$

$$\frac{1}{C_{11}}=\frac{1}{C_1}+\frac{1}{C_3} \qquad \frac{1}{C_{22}}=\frac{1}{C_2}+\frac{1}{C_3} \qquad \frac{1}{C_{12}}=\frac{1}{C_{21}}=-\frac{1}{C_3}.$$

It should be observed that the signs of the mutual coefficients R_{12}, L_{12}, C_{12} are a matter of convention. For example if the conventional directions of I_2 and E_2 are reversed, the signs of the mutual coefficients are reversed.



Sketch 1



Sketch 2

The system of equations (1) possesses two important properties which are largely responsible for the relative simplicity of classical electric circuit theory. First, the equations are linear in both currents and applied electromotive forces. Secondly, the coefficients L_{jk}, R_{jk}, C_{jk} are constants. Important electrotechnical problems exist, in which these properties no longer obtain. The solution, however, for the restricted system of linear equations with constant coefficients is fundamental and its solution can be extended to important problems involving non-linear relations and variable coefficients. These extensions will be taken up briefly in a later chapter.

Another important property is the reciprocal relation among the coefficients; that is $L_{jk}=L_{kj}$: $R_{jk}=R_{kj}$, and $C_{jk}=C_{kj}$. It is easily shown that these reciprocal relations mean that there are no concealed sources or sinks of energy. Again important cases exist where the reciprocal relations do not hold. Such exceptions, however, while of physical interest do not affect the mathematical methods of solution, to which the reciprocal relation is not essential.

Returning to equation (1) we shall now derive the *equation of activity*. Multiply the first equation by I_1, the second by I_2, etc. and add: we get

$$\frac{d}{dt}\sum\sum\frac{1}{2}L_{jk}I_jI_k+\frac{d}{dt}\sum\sum\frac{1}{2}\frac{1}{C_{jk}}Q_jQ_k+\sum\sum R_{jk}I_jI_k=\sum E_jI_j. \quad (2)$$

The right hand side is the rate at which the applied forces are supplying energy to the network. The first term on the left is the rate of increase of the magnetic energy

$$\frac{1}{2}\sum\sum L_{jk}I_jI_k,$$

while the second term is the rate of increase of the electric energy

$$\frac{1}{2}\sum\sum\frac{1}{C_{jk}}Q_jQ_k.$$

The last term, $\sum\sum R_{jk}\,I_j\,I_k$, is the rate at which electromagnetic energy is being converted into heat in the network. Consequently in the electrical network, the magnetic energy is a homogeneous quadratic function of the currents, the electric energy is a homogeneous quadratic function of the charges, and the rate of dissipation is a homogeneous quadratic function of the currents. In Maxwell's dynamical theory of electrical networks, these relations were written down at the start and the circuit equations then derived by an application of Lagrange's dynamic equations to the homogeneous quadratic functions.

Returning to equations (1), we observe that, due to the presence of the integral sign, they are integro-differential equations. They are, however, at once reducible to differential equations by the substitution $I = dQ/dt$, whence they become

$$\left(L_{11}\frac{d^2}{dt^2}+R_{11}\frac{d}{dt}+S_{11}\right)Q_1+\ldots+\left(L_{1n}\frac{d^2}{dt^2}+R_{1n}\frac{d}{dt}+S_{1n}\right)Q_n=E_1,$$

$$\text{- -} \qquad (3)$$

$$\left(L_{n1}\frac{d^2}{dt^2}+R_{n1}\frac{d}{dt}+S_{n1}\right)Q_1+\ldots+\left(L_{nn}\frac{d^2}{dt^2}+R_{nn}\frac{d}{dt}+S_{nn}\right)Q_n=E_n.$$

Here, as a matter of convenience, we have written $1/C_{jk}=S_{jk}$. It is often more convenient, at least at the outset, to deal with equations (3) rather than (1).

The Exponential Solution

In taking up the mathematical solution of equations (1), we shall start with the *exponential solution*. This is of fundamental importance, both theoretically and practically. It serves as the most direct introduction to the Heaviside Operational Calculus, and in addition furnishes the basis of the *steady-state* solution, or the theory of alternating currents.

To derive this solution we set $E_1=F_1e^{\lambda t}$ and put all the other forces $E_2, \ldots E_n$ equal to zero. This latter restriction is a mere matter of convenience, and, in virtue of the linear character of the equations, involves no loss of generality.

Now, corresponding to $E_1=F_1e^{\lambda t}$, let us assume a solution of the form

$$I_j=J_je^{\lambda t} \qquad (j=1,2\ldots n)$$

where J_j is a constant. So far this is a pure assumption, and its correctness must be verified by substitution in the differential equations.

Now if $I_j=J_je^{\lambda t}$, it follows at once that

$$\frac{d}{dt}I_j=\lambda I_j=\lambda J_je^{\lambda t}$$

and

$$\int I_jdt=\frac{1}{\lambda}I_j=\frac{1}{\lambda}J_je^{\lambda t}.$$

Now substitute these relations in equations (1) and cancel the common factor $e^{\lambda t}$. We then get the system of simultaneous equations

$$(\lambda L_{11}+R_{11}+1/\lambda C_{11})J_1+\ldots+(\lambda L_{1n}+R_{1n}+1/\lambda C_{1n})J_n=F_1,$$
$$(\lambda L_{21}+R_{21}+1/\lambda C_{21})J_1+\ldots+(\lambda L_{2n}+R_{2n}+1/\lambda C_{2n})J_n=o,$$
$$\text{- -} \qquad (4)$$
$$(\lambda L_{n1}+R_{n1}+1/\lambda C_{n1})J_1+\ldots+(\lambda L_{nn}+R_{nn}+1/\lambda C_{nn})J_n=o.$$

We note that this is a system of simultaneous *algebraic* equations from which the time factor has disappeared. It is this that makes the exponential solution so simple, since we can immediately pass from differential equations to algebraic equations. In these algebraic equations, n in number, there are n unknown quantities $J_1, \ldots J_n$. These can therefore all be uniquely determined. We thus see that the assumed form of solution is possible.

The notation of equations (4) may be profitably simplified as follows: write

$$\lambda L_{jk}+R_{jk}+1/\lambda C_{jk}=z_{jk}(\lambda)=z_{jk}$$

and we have

$$\begin{aligned} z_{11}J_1+z_{12}J_2+\ldots+z_{1n}J_n&=F_1,\\ z_{21}J_1+z_{22}J_2+\ldots+z_{2n}J_n&=o,\\ &\text{- - - - - - - - - -}\\ z_{n1}J_1+z_{n2}J_2+\ldots+z_{nn}J_n&=o. \end{aligned} \tag{5}$$

The solution of this system of equations is

$$J_j=\frac{M_{j1}(\lambda)}{D(\lambda)}F_1=\frac{M_{j1}}{D}F_1$$

and

$$I_j=\frac{M_{j1}}{D}F_1e^{\lambda t}=\frac{F_1}{Z_{j1}}e^{\lambda t} \tag{6}$$

where D is the determinant of the coefficients,

$$\begin{vmatrix} z_{11} & z_{12} & z_{13} & \ldots & z_{1n}\\ z_{21} & z_{22} & z_{23} & \ldots & z_{2n}\\ z_{31} & z_{32} & \ldots & \ldots & z_{3n}\\ - & - & - & - & -\\ z_{n1} & z_{n2} & \ldots & \ldots & z_{nn} \end{vmatrix} \tag{7}$$

and M_{j1} is the cofactor, or minor with proper sign, of the jth column and first row.

I shall not attempt to discuss the theory of determinants on which this solution is based.[1] We may note, however, one important property. Since $z_{jk}=z_{kj}$, $M_{jk}=M_{kj}$. From this the Reciprocal Theorem follows immediately. This may be stated as follows:

If a force $Fe^{\lambda t}$ is applied in the jth mesh, or branch, of the network, the current in the kth mesh, or branch, is by the foregoing

$$\frac{M_{kj}}{D}Fe^{\lambda t}.$$

Now apply the same force in the kth mesh, or branch, then the current in the jth mesh is

$$\frac{M_{jk}}{D}Fe^{\lambda t}.$$

[1] For a remarkably concise and complete discussion of the exponential solution by aid of the theory of determinants, see Cisoidal Oscillations, Trans. A. I. E. E., 1911, by G. A. Campbell.

Comparing these expressions and remembering that $M_{kj} = M_{jk}$, it follows that the current in the kth branch corresponding to an exponential impressed e.m.f. in the jth branch, is equal to the current in the jth branch corresponding to the same e.m.f. in the kth branch. This relation is of the greatest technical importance.

In many important technical problems we are interested only in two accessible branches, such as the sending and receiving. In such cases, where we are not concerned with the currents in the other meshes or branches, it is often convenient to eliminate them from the equation. Thus suppose that we have electromotive forces E_1 and E_2 in meshes 1 and 2 and are concerned only with the currents in these meshes. If we solve equations 3, 4, . . n, $n-2$ in number, for I_3 . . I_n in terms of I_1 and I_2 and then substitute in (1) and (2) we get

$$\begin{aligned} Z_{11}I_1 + Z_{12}I_2 &= E_1 \\ Z_{21}I_1 + Z_{22}I_2 &= E_2. \end{aligned} \tag{8}$$

The Steady State Solutions

The steady state solution, on which the whole theory of alternating currents depends, is immediately derivable from the exponential solution. Let us suppose that $E_2 = E_3 = \ldots = E_n = o$ and that $E_1 = F\cos(\omega t - \theta)$. Now by virtue of the well known formula in the theory of the complex variable, $\cos x = \frac{1}{2}e^{ix} + \frac{1}{2}e^{-ix}$, we can write

$$\begin{aligned} E_1 &= \tfrac{1}{2}Fe^{i(\omega t-\theta)} + \tfrac{1}{2}Fe^{-i(\omega t-\theta)}, \\ &= \tfrac{1}{2}(\cos\theta - i\sin\theta)Fe^{i\omega t} + \tfrac{1}{2}(\cos\theta + i\sin\theta)Fe^{-i\omega t}, \\ &= \tfrac{1}{2}F'e^{i\omega t} + \tfrac{1}{2}F''e^{-i\omega t}. \end{aligned} \tag{9}$$

Now, by virtue of this formula, the applied electromotive force E_1 consists of two exponential forces, one varying as $e^{i\omega t}$ and the other as $e^{-i\omega t}$. Hence it is easy to see that the currents are made up of two components, thus

$$I_j = J_j'e^{i\omega t} + J_j''e^{-i\omega t} \quad (j = 1, 2 \ldots n) \tag{10}$$

and we have merely to use the exponential solution given above, substituting for λ, $i\omega$ and $-i\omega$ respectively. That is,

$$J_j' = \frac{1}{2}\frac{F'}{Z_{j1}(i\omega)} \text{ and } J_j'' = \frac{1}{2}\frac{F''}{Z_{j1}(-i\omega)}$$

or

$$I_j = \frac{1}{2}\frac{Fe^{-i\theta}}{Z_{j1}(i\omega)}e^{i\omega t} + \frac{1}{2}\frac{Fe^{i\theta}}{Z_{j1}(-i\omega)}e^{-i\omega t}.$$

The second term is the conjugate imaginary of the first, so that

$$\begin{aligned} I_j &= R\frac{Fe^{-i\theta}}{Z_{j1}(i\omega)}e^{i\omega t} \\ &= R\frac{F}{Z_{j1}(i\omega)}e^{i(\omega t-\theta)} \\ &= R\frac{F}{|Z_{j1}(i\omega)|}e^{i(\omega t-\theta-\phi)} \\ &= \frac{F}{|Z(i\omega)|}\cos(\omega t-\theta-\phi). \end{aligned}$$

We thus arrive at the rule for the steady state solution:

If the applied e.m.f. is $F\cos(\omega t-\theta)$, substitute $i\omega$ for d/dt in the differential equations, determine the impedance function

$$Z(i\omega)=D(i\omega)/M(i\omega) \tag{11}$$

by the solution of the algebraic equations, and write it in the form

$$Z(i\omega)=|Z(i\omega)|e^{i\phi}. \tag{12}$$

Then the required solution is

$$I=\frac{F}{|Z(i\omega)|}\cos(\omega t-\theta-\phi). \tag{13}$$

This in compact form contains the whole theory of the symbolic solution of alternating current problems.

The Complementary Solution

So far in the solutions which we have discussed the currents are of the same type as the impressed forces: that is to say in physical language, the currents are "forced" currents and vary with time in precisely the same manner as do the electromotive forces. Such currents are, however, in general only part of the total currents. In addition to the forced currents we have also the characteristic oscillations; or, in mathematical language, the complete solution must include both particular and complementary solutions. This may be shown as follows: Let $I_1', \ldots. I_n'$ be solutions of the complementary equations,

$$\begin{aligned} &\left(L_{11}\frac{d}{dt}+R_{11}+\frac{1}{C_{11}}\int dt\right)I_1'+\ldots+\left(L_{1n}\frac{d}{dt}+R_{1n}+\frac{1}{C_{1n}}\int dt\right)I_n'=o, \\ &-\;- \\ &\left(L_{n1}\frac{d}{dt}+R_{n1}+\frac{1}{C_{n1}}\int dt\right)I_1'+\ldots+\left(L_{nn}\frac{d}{dt}+R_{nn}+\frac{1}{C_{nn}}\int dt\right)I_n'=o. \end{aligned} \tag{14}$$

Then if $I_1 \ldots I_n$ is a solution of (1), $I_1+I_1', \ldots I_n+I_n'$, is also a solution.

To derive the solution of the complementary system of equations (11), assume that a solution exists of the form

$$I_j'=J_j' e^{\lambda t} \qquad (j=1,2 \,..\, n)$$

so that $d/dt=\lambda$ and $\int dt=1/\lambda$. Substitute in equations (14) and cancel out the common factor $e^{\lambda t}$. Then we have

$$\begin{array}{c} Z_{11}(\lambda)J_j'+ \,\ldots\, +Z_{1n}(\lambda)J_n'=o, \\ \text{- -} \\ Z_{n1}(\lambda)J_1'+ \,\ldots\, +Z_{nn}(\lambda)J_n'=o. \end{array} \tag{15}$$

This is a system of n homogeneous equations in the unknown quantities $J_1', \ldots J_n'$. The condition that a finite solution shall exist is that, in accordance with a well known principle of the theory of equations, the determinant of the coefficients shall vanish. That is,

$$D(\lambda)=\begin{vmatrix} Z_{11}(\lambda) \,\ldots\, Z_{1n}(\lambda) \\ \text{- - - - - - - - - - - -} \\ \text{- - - - - - - - - - - -} \\ Z_{n1}(\lambda) \,\ldots\, Z_{nn}(\lambda) \end{vmatrix}=o. \tag{16}$$

Consequently the possible values of λ must be such that this equation is satisfied. In other words, λ must be a root of the equation $D(\lambda)=o$. Let these roots be denoted by $\lambda_1, \lambda_2 \,..\, \lambda_m$. Then, assigning to λ any one of these values, we can determine the ratio J_j'/J_k' from any $(n-1)$ of the equations. That is to say, if we take

$$I_1'=C_1e^{\lambda_1 t}+C_2e^{\lambda_2 t}+ \,\ldots\, +C_me^{\lambda_m t}, \tag{17}$$

substitution in any $(n-1)$ of the equations determines $I_2', \ldots I_n'$. The m constants $C_1, \ldots C_m$ are so far, however, entirely arbitrary, and are at our disposal to satisfy imposed *boundary conditions.*

This introduces us to the idea of boundary conditions which is of the greatest importance in circuit theory. In physical language the boundary conditions denote the state of the system when the electromotive force is applied or when any change in the circuit constants occurs. The number of independent boundary conditions which can, in general, be satisfied is equal to the number of roots of the equation $D(\lambda)=o$. Evidently, therefore, it is physically impossible to impose more boundary conditions than this. On the other hand, if this number of boundary conditions is not specified, the complete solution is indeterminate: That is to say, the problem is not correctly set. As an example of boundary conditions, we may specify that the electromotive force is applied at time $t=o$, and that at this time all the currents in the inductances and all the charges on the condensers are zero.

So far we have been following the classical theory of linear differential equations. We have seen that the forced exponential solution and the derived steady state solution are extremely simple and are mere matters of elementary algebra. The practical difficulties in the classical method of solutions begin with the determination of the constants $C_1, \ldots C_m$ of the complementary solution as well as the roots $\lambda_1, \ldots \lambda_m$ of the equation $D(\lambda)=o$. It is at this point that Heaviside broke with classical methods, and by considering special boundary conditions of great physical importance, and particular types of impressed forces, laid the foundations of original and powerful methods of solution. We shall therefore at this point follow Heaviside's example and attack the problem from a different standpoint. In doing this we shall not at once take up an exposition of Heaviside's own method of attack. We shall first establish some fundamental theorems which are extremely powerful and will serve us as a guide in interpreting and rationalizing the Heaviside Operational Calculus.

CHAPTER II

The Solution when an Arbitrary Force is Applied to the Network in a State of Equilibrium

In engineering applications of electric circuit theory there are three outstanding problems:

(1) The steady state distribution of currents and potentials when the network is energized by a sinusoidal electromotive force. This problem is the subject of the theory of alternating currents which forms the basis of our calculations of power lines and the more elaborate networks of communication systems.

(2) The distribution of currents and potentials in the network in response to an arbitrary electromotive force applied to the network in a state of equilibrium, i.e., applied when the currents and charges in the network are identically zero.

(3) The effect on the distribution of currents and potentials of suddenly changing a circuit constant or connection, such as opening or closing a switch, while the system is energized.

We shall base our further analysis of circuit theory on the solutions of problem (2), for the following reasons:

(A) It is essentially a generalization of the Heaviside problem and its solution will furnish us a key to the correct understanding and interpretation of operational methods and lead to an auxiliary formula from which the rules of the Operational Calculus are directly deducible.

(B) The solution of problem (2) carries with it the solution of problem (3) and also serves as a basis for the theory of alternating currents.

(C) The solution of problem (2) leads directly to an extension of circuit theory to the case where the network contains variable elements: i.e., circuit elements which vary with time and in which non-linear relations obtain.

Problem (2) is therefore the fundamental problem of circuit theory and the formula which we shall now derive may be termed the fundamental formula of circuit theory.

Consider a network in any branch of which, say branch 1, a unit e.m.f. is inserted at time $t=o$, the network having been previously in equilibrium. By unit e.m.f. is meant an electromotive force which has the value unity for all positive values of time ($t \geqq o$). Let the resultant current in any branch, say branch n, be denoted by $A_{n1}(t)$. $A_{n1}(t)$ will be termed the *indicial admittance* of branch n with respect to branch 1—or, more fully, the transfer indicial admittance.

The indicial admittance, aside from its direct physical significance, plays a fundamental role in the mathematical theory of electric circuits. In words, it may be defined as follows: The indicial admittance, $A_{n1}(t)$, is equal to the ratio of the current in branch n, expressed as a time function, to the magnitude of the steady e.m.f. suddenly inserted at time $t=o$ in branch 1. It is evidently a function which is zero for negative values of time and approaches either zero or a steady value (the d.c. admittance) for all actual dissipative systems, as t approaches infinity. It may be noted that, aside from its mathematical determination, which will engage our attention later, it is an experimentally determinable function.

We note, in passing, an important property of the indicial admittance $A_{jk}(t)$, which is deducible from the reciprocal theorem:[2] this is that $A_{jk}(t)=A_{kj}(t)$. That is to say, the value of the transfer indicial admittance is unchanged by an interchange of the driving point and receiving point. It is therefore immaterial in the expression $A_{jk}(t)$ whether the e.m.f. is inserted in branch j and the current measured in branch k, or vice-versa. In general, unless we are concerned with particular branches, the subscripts will be omitted and we shall simply write $A(t)$, it being understood that any two branches

[2] Exceptions to this relation exist where the network contains sources of energy such as amplifiers. These need not engage our attention here.

or a single branch (for the case of equal subscripts) may be under consideration.

From the linear character of the network, it is evident that if a steady e.m.f. $E=E_\tau$ is inserted at time $t=\tau$, the network being in equilibrium, the resultant current is

$$E_\tau \cdot A(t-\tau).$$

Generalizing still further, suppose that steady e.m.fs. $E_o, E_1, E_2, \ldots E_n$ are impressed in the same branch at the respective times $\tau_o, \tau_1, \tau_2 \ldots \tau_n$; the resultant current is evidently

$$E_oA(t)+E_1A(t-\tau_1)+\ldots+E_nA(t-\tau_n)=\sum_{j=o}^{n}E_jA(t-\tau_j). \tag{18}$$

To apply the foregoing to our problem we suppose that there is applied to the network, initially in a state of equilibrium, an e.m.f. $E(t)$ which has the following properties.

1. It is identically zero for $t<o$.
2. It has the value $E(o)$ for $o\leq t\leq \Delta t$.
3. It has the value $E(o)+\Delta_1E$ for $\Delta t\leq t<2\Delta t$.
4. It has the value $E(o)+\Delta_1E+\Delta_2E$ for $2\Delta t\leq t<3\Delta t$.

In other words it has the increment Δ_jE at time $t=j\Delta t$.

Evidently then the resultant current $I(t)$ is

$$E_oA(t)+\Delta_1EA(t-\Delta t)+\ldots+\Delta_nE.A(t-n\Delta t).$$

Now evidently if the interval Δt is made shorter and shorter, then in the limit $\Delta t\rightarrow dt$ and $j\Delta t=\tau$ and

$$\Delta_jE=\frac{d}{d\tau}E(\tau)d\tau.$$

Passing to the limit in the usual manner this summation becomes a definite integral and we get

$$I(t)=E(o)A(t)+\int_0^t A(t-\tau)\frac{d}{d\tau}E(\tau)d\tau. \tag{19}$$

Finally by obvious transformations of the expression we arrive at the fundamental formula of circuit theory

$$I(t)=\frac{d}{dt}\int_0^t A(t-\tau)E(\tau)d\tau, \tag{20}$$

$$=\frac{d}{dt}\int_0^t E(t-\tau)A(\tau)d\tau. \tag{20-a}$$

For completeness we write down the following equivalents of (20) and (20-a)

$$I(t)=A(o)E(t)+\int_0^t A'(t-\tau)E(\tau)d\tau, \qquad (20\text{-b})$$

$$=A(o)E(t)+\int_0^t A'(\tau)E(t-\tau)d\tau, \qquad (20\text{-c})$$

$$=E(o)A(t)+\int_0^t E'(t-\tau)A(\tau)d\tau, \qquad (20\text{-d})$$

$$=E(o)A(t)+\int_0^t E'(\tau)A(t-\tau)d\tau. \qquad (20\text{-e})$$

where the primes denote differentiation with respect to the argument. Thus $A'(t)=d/dt\,A(t)$.

These equations are the fundamental formulas which mathematically relate the current to the type of applied electromotive force and the constants and connections of the system, and constitute the first part of the solution of our problem. The most important immediate deductions from these formulas are expressed in the following theorems.

1. The indicial admittance of an electrical network completely determines, within a single integration, the behavior of the network to all types of applied electromotive forces. As a corollary, a knowledge of the indicial admittance is the sole information necessary to completely predict the performance and characteristics of the system, including the steady state.

2. The applied e.m.f. and the inidical admittance are similarly and coequally related to the resultant current in the network. As a corollary the form of the current may be modified either by changing the constants and connections of the network or by modifying the form of the applied e.m.f.

3. Since the applied e.m.f. may be discontinuous these formulas determine not only the building up of the current in response to an applied e.m.f. but also its subsidence to equilibrium when the e.m.f. is removed and the network left to itself. In brief, formulas (20) reduce the whole problem to a determination of the indicial admittance of the network. In addition, as we shall see, they lead directly to an integral equation which determines this function.

It is of interest to show the relation between formulas (20) and the usual steady state equations. To do this let the e.m.f., applied at time $t=o$, be $E \sin(\omega t+\theta)$. Substitution in formula (20-b) and rearrangement gives

$$I(t)=A(o)E\sin(\omega t+\theta)$$
$$+E\sin(\omega t+\theta)\int_0^t \cos\omega\tau A'(\tau)d\tau$$
$$-E\cos(\omega t+\theta)\int_0^t \sin\omega\tau A'(\tau)d\tau \qquad (21)$$

where $\qquad A'(t)=\frac{d}{dt}A(t).$

Now this can be resolved into two parts

$$E\sin(\omega t+\theta)\left\{A(o)+\int_0^\infty \cos\omega\tau A'(\tau)d\tau\right\}$$
$$-E\cos(\omega t+\theta)\left\{\int_0^\infty \sin\omega\tau A'(\tau)d\tau\right\} \qquad (22)$$

which is the *final steady state*, and

$$-E\sin(\omega t+\theta)\int_t^\infty \cos\omega\tau A'(\tau)d\tau$$
$$+E\cos(\omega t+\theta)\int_t^\infty \sin\omega\tau A'(\tau)d\tau \qquad (23)$$

which is the *transient distortion*, which ultimately dies away for sufficiently large values of time.

To correlate the foregoing expressions for the steady state with the usual formulas we observe that if the symbolic impedance of the network at frequency $\omega/2\pi$ be denoted by $Z(i\omega)$, and if we write

$$\frac{1}{Z(i\omega)}=\alpha(\omega)+i\beta(\omega)$$

then the steady state current is

$$E[\alpha(\omega).\sin(\omega t+\theta)+\beta(\omega).\cos(\omega t+\theta)].$$

Comparison with (22) gives at once

$$\alpha(\omega)=A(o)+\int_0^\infty \cos\omega\tau\, A'(\tau)d\tau, \qquad (24)$$

$$\beta(\omega)=-\int_0^\infty \sin\omega\tau\, A'(\tau)d\tau. \qquad (25)$$

The Integral Equation for the Indicial Admittance

So far we have tacitly assumed that the indicial admittance is known. As a matter of fact its determination constitutes the essential part of our problem. It is, in fact, the Heaviside problem, and its investigation, to which we now proceed, will lead us directly to the Operational Calculus.

Heaviside's method in investigating this problem was intuitive and "experimental". We, however, shall establish a very general integral equation from which we shall directly deduce his methods and extensions thereof.

Let us suppose that an e.m.f. e^{pt}, where p is either positive real quantity or complex with real part positive, is suddenly impressed on the network at time $t=o$. It follows from the foregoing theory that the resultant current $I(t)$ will be made up of two parts, (1) a forced exponential part which varies with time as e^{pt}, and (2) a complementary part which we shall denote by $y(t)$. The exponential or "forced" component is simply e^{pt} $Z(p)$, where $Z(p)$ is functionally of the same form as the usual symbolic or complex impedance $Z(i\omega)$. It is gotten from the differential equations of the problem, as explained in a preceding section, by replacing d^n dt^n by p^n, cancelling out the common factor e^{pt}, and solving the resulting algebraic equation. The complementary or characteristic component, denoted by $y(t)$, depends on the constants and connections of the network, and on the value of p. It does not, however, contain the factor e^{pt} and it dies away for sufficiently large value of t, in all actual dissipative systems. Thus

$$I(t)=\frac{e^{pt}}{Z(p)}+y(t). \qquad (26)$$

Now return to formula (20-a) and replace $E(t)$ by e^{pt}. We get

$$I(t)=\frac{d}{dt}e^{pt}\int_0^t A(\tau)e^{-p\tau}d\tau$$

which can be written as

$$\frac{d}{dt}\left\{e^{pt}\int_0^\infty A(\tau)e^{-p\tau}d\tau-e^{pt}\int_t^\infty A(\tau)e^{-p\tau}d\tau\right\}.$$

Carrying out the indicated differentiation this becomes

$$I(t)=pe^{pt}\int_0^\infty A(\tau)e^{-p\tau}d\tau-pe^{pt}\int_t^\infty A(\tau)e^{-p\tau}d\tau+A(t). \qquad (27)$$

Equating the two expressions (26) and (27) for $I(t)$ and dividing through by e^{pt} we get

$$\frac{1}{Z(p)}+y(t)e^{-pt}=p\int_0^\infty A(\tau)e^{-p\tau}d\tau-p\int_t^\infty A(\tau)e^{-p\tau}d\tau+A(t)e^{-pt}. \quad (28)$$

This equation is valid for all values of t. Consequently if we set $t=\infty$, and if the real part of p is positive, only the first term on the right and the left hand side of the equation remain, the rest vanishes, and we get

$$\frac{1}{pZ(p)}=\int_0^\infty A(t)e^{-pt}dt. \quad (29)$$

This is an integral equation [3] *valid for all positive real values of* p, *which completely determines the indicial admittance* $A(t)$. It is on this equation that we shall base our discussion of operational methods and from which we shall derive the rules of the Operational Calculus. Equations (20) and (29) constitute a complete mathematical formulation of our problem, and from them the complete solution is obtainable without further recourse to the differential equations, or further consideration of boundary conditions.

To summarize the preceding: we have reduced the determination of the current in a network in response to an electromotive force $E(t)$, impressed on the network at reference time $t=o$, to the mathematical solution of two equations: first the integral equation

$$\frac{1}{pZ(p)}=\int_0^\infty A(t)e^{-pt}dt \quad (29)$$

and second, the definite integral

$$I(t)=\frac{d}{dt}\int_0^t A(t-\tau)E(\tau)d\tau. \quad (20)$$

It will be observed that in deducing these equations we have merely postulated (1) the linear and invariable character of the network and (2) the existence of an exponential solution of the type $e^{pt}/Z(p)$ for positive values of p. Consequently, while we have so far discussed these formulas in terms of the determination of the current in a finite network, they are not limited in their application to this specific problem. In this connection it may be well to call attention explicitly to the following points.

[3] An integral equation is one in which the unknown function appears under the sign of integration. (29) is an integral equation of the Laplace type. If $Z(p)$ is specified, $A(t)$ is uniquely determined. Methods for solving the integral equations are considered in detail later, in connection with the exposition of the Operational Calculus. The phrase "all positive values of p" will be understood as meaning all values of p in the right hand half of the complex plane.

The formulas and methods deduced above apply not only to finite networks, involving a finite system of linear equations, but to infinite networks and to transmission lines, involving infinite systems of equations, and partial differential equations: in fact to all electrical and dynamical systems in which the connections and constants are linear and invariable.

Secondly the variable determined by formula (20) and (29) need not, of course, be the current. It may equally well be the charge, potential drop, or any of the variables with which we may happen to be concerned. This fact may be explicitly recognized by writing the formulas as:

$$\frac{1}{pH(p)} = \int_0^\infty h(t)e^{-pt}dt, \tag{30}$$

$$x(t) = \frac{d}{dt}\int_0^t h(t-\tau)E(\tau)d\tau. \tag{31}$$

Here $E(t)$ is the applied e.m.f., $x(t)$ is the variable which we desire to determine (charge, current, potential drop, etc.), and

$$x = E/H(p) \tag{32}$$

is the operational equation. $H(p)$ therefore corresponds to and is determined in precisely the same way as the impedance $Z(p)$, but it may not have the physical significance or the dimensions of an impedance. Similarly in character and function, $h(t)$ corresponds to the indicial admittance, though it may not have the same physical significance. It is a generalization of the indicial admittance and may be appropriately termed the *Heaviside Function.* Similarly $H(p)$ may be termed the *generalized impedance function.*

CHAPTER III

The Heaviside Problem and the Operational Equation

The physical problem which Heaviside attacked and which led to his Operational Calculus was the determination of the response of a network or electrical system to a "unit e.m.f." (zero before, unity after time $t=o$) with, of course, the understanding that the system is in equilibrium when the electromotive force is applied. His problem is therefore, essentially that of the determination of the indicial admittance. In our exposition and critique of Heaviside's method of dealing with this problem we shall accompany an account of his own method of solution with a parallel solution from the corresponding integral equation of the problem.

Heaviside's first step in attacking this problem was to start with the differential equations, and replace the differential operator d/dt by the symbol p, and the operation $\int dt$ by $1/p$, thus reducing the equations to an algebraic form. He then wrote the impressed e.m.f. as 1 (unity), thus limiting the validity of the equations to values of $t \geqq o$. The formal solution of the algebraic equations is straightforward and will be written as

$$h = 1/H(p) \tag{33}$$

where h is the "generalized indicial admittance," or Heaviside function (denoting current, charge, potential or any variable with which we are concerned) and $H(p)$ is the corresponding generalized impedance. Thus, if we are concerned with the current in any part of the network, we write

$$A = 1/Z(p). \tag{34}$$

The more general notation is desirable, however, as indicating the wider applicability of the equation.

The equations

$$h = 1/H(p)$$

$$A = 1/Z(p)$$

are the *Heaviside Operational Equations*. They are, as yet, purely symbolic and we have still the problem of determining their explicit meaning and in particular the significance of the operator p.

Comparison of the Heaviside Operational Equations with the integral equations (29) and (30) of the preceding chapter leads to the following fundamental theorem.

The Heaviside Operational Equations

$$A = 1/Z(p)$$

$$h = 1/H(p)$$

are merely the symbolic or short-hand equivalents of the corresponding integral equations

$$\frac{1}{pZ(p)} = \int_0^\infty A(t)e^{-pt}dt$$

$$\frac{1}{pH(p)} = \int_0^\infty h(t)e^{-pt}dt.$$

The integral equations, therefore, supply us with the meaning and significance of the operational equations, and from them the rules of the Operational Calculus are deducible.

By virtue of this theorem, we have the advantage, at the outset, of a key to the meaning of Heaviside's operational equations, and a means of checking and deducing his rules of solution. This will serve us as a guide throughout our further study.

Returning now to Heaviside's own point of view and method of attack, his reasoning may be described somewhat as follows: The operational equation

$$h = 1/H(p)$$

is the full equivalent of the differential equations of the problem and must therefore contain the information necessary to the solution provided we can determine the significance of the symbolic operator p. The only way of doing this, when starting with the operational equation, is one of induction: that is, we must compare the operational equation with known solutions of specific problems and thus attempt to infer by induction general rules for interpreting the operational equation and converting it into the required explicit solution.

The Power Series Solution

Let us start with the simplest possible problem: the current in response to a "unit e.m.f." in a circuit consisting of an inductance L in series with a resistance R.

The differential equation of the problem is

$$L\frac{d}{dt}A + RA = 1, \qquad t \geqq o,$$

where A is the indicial admittance. Consequently replacing d/dt by p, the operational equation is

$$A = \frac{1}{pL+R}.$$

The explicit solution is easily derived: it is

$$A = \frac{1}{R}(1 - e^{-\alpha t})$$

where $\alpha = R/L$. Note that this makes the current initially zero, so that the equilibrium boundary condition at $t = o$ is satisfied.

Now suppose that we expand the operational equation in inverse powers of p: we get, formally,

$$A = \frac{1}{pL}\frac{1}{1+\alpha/p} = \frac{1}{R}\frac{\alpha}{p}\frac{1}{1+\alpha/p} = \frac{1}{R}\left\{\frac{\alpha}{p} - \left(\frac{\alpha}{p}\right)^2 + \left(\frac{\alpha}{p}\right)^3 - \left(\frac{\alpha}{p}\right)^4 + \dots\right\}$$

by the Binomial Theorem.

Now expand the explicit solution as a power series in t: it is

$$A=\frac{1}{R}\left\{\frac{\alpha t}{1!}-\frac{(\alpha t)^2}{2!}+\frac{(\alpha t)^3}{3!}-\ldots\ldots\right\}.$$

Comparing the two expansions we see at once that the operational expansion is converted into the explicit solution by assigning to the symbol $1/p^n$ the value $t^n/n!$. It was from this kind of inductive inference that Heaviside arrived at his power series solution.

Now there are several important features in the foregoing which require comment. In the first place the operational equation is converted into the explicit solution only by a particular kind of expansion, namely an expansion in inverse powers of the operator p. For example, if in the operational equation

$$A=\frac{1}{R}\frac{\alpha/p}{1+\alpha/p}$$

we replace $1/p$ by $t/1!$ we get

$$A=\frac{1}{R}\frac{\alpha t}{1+\alpha t}$$

which is incorrect. Furthermore, if we expand in ascending instead of descending powers of p, namely

$$A=\frac{1}{R}\left\{1-(p/\alpha)+(p/\alpha)^2-\ldots\ldots\right\}$$

no correlation with the explicit solution is possible and no significance can be attached to the expansion. We thus infer the general principle, and we shall find this inference to be correct, that the operational equation is convertible into the explicit solution only by the proper choice of expansion of the impedance function, or rather its reciprocal.

In the second place we notice that in writing down the operational equation and then converting it into the explicit solution no consideration has been given to the question of boundary conditions. This is one of the great advantages of the operational method: the boundary conditions, *provided they are those of equilibrium*, are automatically taken care of. This will be illustrated in the next example:

Let a "unit e.m.f." be impressed on a circuit consisting of resistance R, inductance L, and capacity C: required the resultant charge on the condenser.

The differential equation for the charge Q is

$$\left(L\frac{d^2}{dt^2}+R\frac{d}{dt}+1/C\right)Q=1, \qquad t\geqq o.$$

Consequently the operational formula is

$$Q=\frac{1}{Lp^2+Rp+1/C}$$

$$=\frac{1}{Lp^2}\frac{1}{1+a/p+b/p^2}, \text{ where } a=\frac{R}{L} \text{ and } b=\frac{1}{LC}.$$

This can be expanded by the Binomial Theorem as

$$Q=\frac{1}{Lp^2}\left\{1-\left(\frac{a}{p}+\frac{b}{p^2}\right)+\left(\frac{a}{p}+\frac{b}{p^2}\right)^2-\left(\frac{a}{p}+\frac{b}{p^2}\right)^3+\ldots\right\}.$$

Performing the indicated operations and collecting in inverse powers of p, the first few terms of the expansion are:—

$$\frac{1}{Lp^2}\left\{1-\frac{c_1}{p}-\frac{c_2}{p^2}+\frac{c_3}{p^3}+\frac{c_4}{p^4}-\frac{c_5}{p^5}-\frac{c_6}{p^6}+\ldots\right\}$$

where

$$c_1=a$$

$$c_2=b-a^2$$

$$c_3=2ab-a^3$$

$$c_4=b^2-3a^2b+a^4$$

$$c_5=3ab^2-4a^3b+a^5$$

$$c_6=b^3-6a^2b^2+5a^4b-a^6$$

- - - - - - - - - - - - - - - -

We infer therefore that in accordance with the rule of replacing $1/p^n$ by $t^n/n!$ the solution is:—

$$Q=\frac{1}{L}\left\{\frac{t^2}{2!}-c_1\frac{t^3}{3!}-c_2\frac{t^4}{4!}+c_3\frac{t^5}{5!}+c_4\frac{t^6}{6!}-\ldots\right\}.$$

Owing to the complicated character of the coefficients in the expansion, the series cannot be recognized and summed by inspection. If, however, we put $R=o$) then $a=o$, and the series becomes

$$C\left\{\frac{1}{2!}\left(\frac{t}{\sqrt{LC}}\right)^2-\frac{1}{4!}\left(\frac{t}{\sqrt{LC}}\right)^4+\frac{1}{6!}\left(\frac{t}{\sqrt{LC}}\right)^6-\ldots\right\}$$

whence

$$Q=C\left\{1-\cos\left(t/\sqrt{LC}\right)\right\}.$$

We have still to verify this solution by comparison with the explicit solution of the differential equation. This is of the form

$$Q=C+k_1e^{\lambda_1 t}+k_2e^{\lambda_2 t}$$

where k_1 and k_2 are constants which must be chosen to satisfy the boundary conditions and λ_1, λ_2 are the roots of the equation

$$L\lambda^2+R\lambda+1\ C=o.$$

Now since we have two arbitrary constants we satisfy the equilibrium condition by making Q and dQ/dt zero at $t=o$, whence

$$C+k_1+k_2=o,$$
$$\lambda_1 k_1+\lambda_2 k_2=o,$$

and

$$k_1=\lambda_2 C/(\lambda_1-\lambda_2),$$
$$k_2=\lambda_1 C/(\lambda_2-\lambda_1).$$

We have also

$$\lambda_1=-\frac{a}{2}+\sqrt{\left(\frac{a}{2}\right)^2-b},$$

$$\lambda_2=-\frac{a}{2}-\sqrt{\left(\frac{a}{2}\right)^2-b}.$$

Writing down the power series expansion of

$$Q=C+k_1e^{\lambda_1 t}+k_2e^{\lambda_2 t},$$

then

$$Q=(C+k_1+k_2)+(k_1\lambda_1+k_2\lambda_2)\frac{t}{1!}$$
$$+(k_1\lambda_1^2+k_2\lambda_2^2)\frac{t^2}{2!}+\ \ldots\ .$$

Introducing the values of k_1, k_2, λ_1, λ_2 given above and comparing with the power series derived from the operational solution we see that they are identical term by term.

This example illustrates two facts. First the power series expansions may be complicated, laborious to derive and of such form that they cannot be recognized and summed by inspection. In fact in arbitrary networks of a large number of meshes or degrees of freedom the evaluation of the coefficients of the power series expansion is extremely laborious.

On the other hand, in such cases, the solution by the classical method presents difficulties far more formidable—in fact insuperable difficulties from a practical standpoint. First there is the location of the roots of the function $H(\lambda)$, which in arbitrary networks is a practical impossibility without a prohibitive amount of labor. Secondly there is the determination of the integration constants to satisfy the imposed boundary conditions: a process, which, while theoretically straightforward, is actually in practice extremely laborious and complicated. We note these points in passing; a more complete estimate of the value of the power series solution will be made later.

To summarize the preceding: Heaviside, generalizing from specific examples otherwise solvable, arrived at the following rule:—

Expand the right hand side of the operational equation

$$h = 1/H(p)$$

in inverse powers of p: *thus*

$$h \sim a_0 + a_1/p + a_2/p^2 + \ldots + a_n/p^n + \ldots .$$

and then replace $\frac{1}{p^n}$ by $t^n/n!$. *The operational equation is thereby converted into the explicit power series solution*:—

$$h = a_0 + a_1 t/1! + a_2 t^2/2! + \ldots + a_n t^n/n! + \ldots \qquad (35)$$

As stated above, this rule was arrived at by pure induction and generalization from the known solution of specific problems. It cannot, therefore, theoretically be regarded as satisfactorily established. The rule can, however, be directly deduced from the integral equation

$$\frac{1}{pH(p)} = \int_0^\infty h(t) e^{-pt} dt.$$

To its derivation from this equation we shall now proceed.

First suppose we *assume* that $h(t)$ admits of the power series expansion

$$h_0 + h_1 t/1! + h_2 t^2/2! + \ldots \ldots$$

Substitute this assumed expansion in the integral, and integrate term by term. The right hand side of the integral equation becomes formally

$$h_0/p + h_1/p^2 + h_2/p^3 + \ldots \ldots$$

by virtue of the formula

$$\int_0^\infty \frac{t^n}{n!} e^{-pt} = \frac{1}{p^{n+1}} \text{ for } p > o.$$

Now expand the left hand side of the integral equation asymptotically in inverse process of p: it becomes

$$a_0/p + a_1/p^2 + a_2/p^3 + \ldots \ldots$$

where

$$a_0 + a_1/p + a_2/p^2 + \ldots .$$

is the asymptotic expansion of $1/H(p)$. Comparing the two expansions and making a term by term identification, we see that $h_n = a_n$ and

$$h(t) = a_0 + a_1 t/1! + a_2 t^2/2! + \dots$$

which agrees with the Heaviside formula.

This procedure, however, while giving the correct result has serious defects from a mathematical point of view. For example, the asymtotic expansion of $1/H(p)$ has usually only a limited region of convergence, and it is only in this region that term by term integration is legitimate. Furthermore we have *assumed* the possibility of expanding $h(t)$ in a power series: an assumption to which there are serious theoretical objections, and which, furthermore, is not always justified. A more satisfactory derivation, and one which establishes the condition for the existence of a power series expansion, proceeds as follows:—

Let $1/H(p)$ be a function which admits of the formal asymptotic expansion

$$\sum_o^\infty a_n/p^n$$

and let it include no component which is asymptotically representable by a series all of whose terms are zero, that is a function $\phi(p)$ such that the limit, as $p \to \infty$, of $p^n \phi(p)$ is zero for every value of n. Such a function is e^{-p}. With this restriction understood, start with the integral equation, and integrate by parts: we get

$$\frac{1}{H(p)} = h(o) + \int_0^\infty e^{-pt} h^{(1)}(t) dt$$

where $h^{(n)}(t)$ denotes $d^n/dt^n h(t)$. Now let p approach infinity: in the limit the integral vanishes and by virtue of the asymptotic expansion

$$1/H(p) \sim \sum_o^\infty a_n/p^n, \tag{36}$$

$1/H(p)$ approaches the limit a_o. Consequently

$$h(o) = a_o.$$

Now integrate again by parts: we get

$$p(1/H(p) - a_o) = h^{(1)}(o) + \int_0^\infty e^{-pt} h^{(2)}(t) dt.$$

Again let p approach infinity: in the limit the left hand side of the equation becomes a_1 and we have

$$h^{(1)}(o) = a_1.$$

Proceeding by successive partial integrations we thus establish the general relation

$$h^{(n)}(o) = a_n.$$

But by Taylor's theorem, the power series expansion of $h(t)$ is simply

$$h(t) = h(o) + h^{(1)}(o)t\ \ 1! + h^{(2)}(o)t^2\ \ 2! + \ldots\ldots$$

whence, *assuming the convergence of this expansion*, we get

$$h(t) = a_o + a_1t\ \ 1! + a_2t^2\ \ 2! + \ldots = \sum_o^\infty a_nt^n\ \ n! \qquad (35)$$

which establishes the power series solution. It should be carefully noted, however, that it does not establish the convergence of the power series solution. As a matter of fact, however, I know of no physical problem in which $H(p)$ satisfies the conditions for an asymptotic expansion, where the power series solution is not convergent. On the other hand many physical problems exist, including those relating to transmission lines, where a power series solution is not derivable and does not exist.

The process of expanding the operational equation in such a form as to permit of its being converted into the explicit solution is what Heaviside calls "algebrizing" the equation. In the case of the power series solution the process of algebrizing consists in expanding the reciprocal of the impedance function in an asymptotic series, thus

$$1\ \ H(p) \backsim a_o + a_1\ \ p + a_2\ \ p^2 + \ldots\ldots$$

Regarded as an expansion in the variable p, instead of as a purely symbolic expansion, this series has usually only a limited region of convergence. This fact need not bother us, however, as the series we are really concerned with is

$$a_o + a_1t\ \ 1! + a_2t^2\ \ 2! + \ldots.$$

It is interesting to note in passing that the latter series is what Borel, the French mathematician, calls the *associated function* of the former, and is extensively employed by him in his researches on the summability of divergent series.

The process of "algebrizing," as in the examples discussed above, may often be effected by a straight forward binomial expansion.

In other cases the form of the generalized impedance function $H(p)$ will indicate by inspection the appropriate procedure. A general process, applicable in all cases where a power series exists, is as follows. Write

$$1/H(p) = 1/H\left(\frac{1}{x}\right) = G(x). \tag{36}$$

Now expand $G(x)$ as a Taylor's series: thus formally

$$G(x) = G(o) + G^{(1)}(o)\frac{x}{1!} + G^{(2)}(o)\frac{x^2}{2!} + \ldots$$

where

$$G^{(n)}(o) = \left[\frac{d^n}{dx^n}G(x)\right]_{x=o}. \tag{37}$$

Denote $\frac{G^{(n)}(o)}{n!}$ by a_n, replace x^n by $1/p^n$, and we have

$$G(x) = 1/H(p) = a_o + a_1/p + a_2/p^2 + \ldots$$

This process of "algebrizing" is formally straightforward and always possible. As implied above, however, in many problems much shorter modes of expansion suggest themselves from the form of the function $H(p)$.

We note here, in passing, that the necessary and sufficient conditions for the existence of a power series solution is the possibility of the formal expansion of $G(x)$ as a power series in x.

At this point a brief critical estimate of the scope and value of the power series solution may be in order. As stated above, in a certain important class of problems relating to transmission lines, a power series does not exist, though a closely related series in fractional powers of t may often be derived. Consequently the power series solution is of restricted applicability. Where, however, a power series does exist, in directness and simplicity of derivation it is superior to any other form of solution. Its chief defect, and a very serious defect indeed, is that except where the power series can be recognized and summed, it is usually practically useless for computation and interpretation except for relatively small values of the time t. This disadvantage is inherent and attaches to all power series solutions. For this reason I think Heaviside overestimated the value of power series as practical or working solutions, and that some of his strictures against orthodox mathematicians and their solutions may be justly urged against the power series solution. He was quite right in insisting that a solution must be capable of either interpretation or computation and quite right in ridiculing those formal solutions which actually conceal rather than reveal the significance of the original differential equations of the problem. On the other hand, the following remark of his indicates to me that Heaviside has a quite exaggerated idea of the value and fundamental character of power series in general: "I regret that the result should be so complicated. But the only alternatives are other equivalent infinite series, or else a definite integral which is of no use until it is evaluated, when the result must be the series (135), or an equivalent one." As a matter of fact the properties of most of the important functions of mathematical physics have been investigated and their values computed by methods other than series expansions. I may add that in technical work the power series solution has proved to be of restricted utility, while definite integrals, which Heaviside[4] particularly despised, have proved quite useful.

The Expansion Theorem Solution[5]

We pass now to the consideration of another extremely important form of solution. Heaviside gives this solution without proof: we shall therefore merely state the solution and then derive it from the integral equation.

Given the operational equation

$$h = 1/H(p)$$

which has the significance discussed above: i.e., the response of the network to a "unit e.m.f.". The explicit solution may be written as

$$h = \frac{1}{H(o)} + \sum_{1}^{n} \frac{e^{p_k t}}{p_k H'(p_k)} \tag{38}$$

where $p_1, p_2 \ldots p_n$ are the n roots of the equation

$$H(p) = o$$

and

$$H'(p_k) = \left[\frac{d}{dp} H(p)\right]_{p=p_k}. \tag{39}$$

As remarked above, this solution, referred to by him as *The Expansion Theorem*, was stated by Heaviside without proof; how he arrived at it will probably always remain a matter of conjecture. Its derivation from the integral equation is, however, a relatively simple matter, though in special cases troublesome questions arise.

[4] Vide a remark of his to the effect that some mathematicians took refuge in a definite integral and called that a solution.

[5] This terminology is due to Heaviside. A more appropriate and physically significant expression would be "The Solution in terms of normal or characteristic vibrations."

The derivation of the expansion solution from the integral equation

$$\frac{1}{pH(p)} = \int^{\infty} h(t)e^{-pt}dt$$

follows immediately from the partial fraction expansion

$$\frac{1}{pH(p)} = \frac{1}{pH(o)} + \sum_{j=1}^{n} \frac{1}{(p-p_j)p_jH'(p_j)} \tag{40}$$

where $p_1, p_2 \ldots p_n$ are the roots of the equation $H(p) = o$, and

$$H'(p_j) = \left\{ \frac{d}{dp} H(p) \right\}_{p=p} . \tag{41}$$

Partial fraction expansions of this type are fully discussed in treatises on algebra and the calculus and the conditions for their existence established. Before discussing the restrictions imposed on $H(p)$ by this expansion, we shall first, assuming its existence, derive the expansion theorem solution.

By virtue of (40) the integral equation is

$$\frac{1}{pH(o)} + \sum_{1}^{n} \frac{1}{(p-p_j)p_jH'(p_j)} = \int_0^{\infty} h(t)e^{-pt}dt. \tag{42}$$

The expansion on the left hand side suggests a corresponding expansion on the right hand side: that is, we suppose that

$$h(t) = h_o(t) + h_1(t) + h_2(t) + \ldots + h_n(t) \tag{43}$$

and specify that these component functions shall satisfy the equations

$$\frac{1}{pH(o)} = \int_0^{\infty} h_o(t)e^{-pt}dt \tag{44}$$

$$\frac{1}{(p-p_j)p_jH'(p_j)} = \int_0^{\infty} h_j(t)e^{-pt}dt \qquad j = 1,2 \ldots n. \tag{45}$$

It follows at once from (43) and direct addition of equations (44) and (45) that (42) is satisfied and hence is solved provided $h_o, \ldots h_n$ can be evaluated from (44) and (45).

Now since

$$\int_0^{\infty} e^{\lambda t}e^{-pt}dt = \frac{1}{p-\lambda} \tag{46}$$

provided the real part of λ is not positive (a condition satisfied in all network problems), we see at once that equations (42) and (43) are satisfied by taking

$$h_o(t) = h_o = \frac{1}{H(o)}, \tag{47}$$

$$h_j(t) = \frac{e^{p_j t}}{p_j H'(p_j)}, \quad j = 1, 2, \ldots n.$$

Consequently from (43) and (47) it follows that

$$h(t) = \frac{1}{H(o)} + \sum_1^n \frac{e^{p_j t}}{p_j H'(p_j)} \tag{48}$$

which establishes the Expansion Theorem Solution.

As implied above, the partial fraction expansion (40), on which the expansion theorem solution depends, imposes certain restrictions on the impedance function $H(p)$. Among these are that $H(p)$ must have no zero root, no repeated roots, and $1/H(p)$ must be a proper fraction. In all finite networks these conditions are satisfied, or by a slight modification, the operational equation can be reduced to the required form. The case of repeated roots, which may occur where the network involves a unilateral source of energy such as an amplifier, can be dealt with by assuming unequal roots and then letting the roots approach equality as a limit. Without entering upon these questions in detail, however, we can very simply and directly establish the proposition that the expansion theorem gives the solution whenever a solution in terms of normal or characteristic vibrations exists. The proof of this proposition proceeds as follows.

It is known from the elementary theory of linear differential equations that the general solution of the set of differential equations, of which the operational equation is $h = 1/H(p)$, is of the form

$$h(t) = C_o + \sum_1^n C_j e^{p_j t}$$

where p_j is the jth root of $H(p) = o$, and $C_o, C_1 \ldots C_n$ are constants of integration which must be so chosen as to satisfy the system of differential equations and the imposed boundary conditions. The summation is extended over all the roots of $H(p)$ which is supposed not to have a zero root or repeated roots.

Now substitute this known form of solution in the integral equation of the problem and carry out the integration term by term. We get

$$\frac{1}{H(p)}=C_o+p\sum\frac{C_j}{p-p_j}. \tag{49}$$

Setting $p=o$, we have at once

$$C_o=1/H(o). \tag{50}$$

To determine C_j let $p=p_j+q$ where q is a small quantity ultimately to be set equal to zero, and write the equation as

$$C_oH(p)+\sum\frac{pH(p)}{p-p_j}C_j=1. \tag{51}$$

If now $p=p_j+q$ and q approaches zero, this becomes in the limit

$$p_jH'(p_j)C_j=1 \tag{52}$$

or

$$C_j=\frac{1}{p_jH'(p_j)}, \tag{53}$$

whence

$$h(t)=\frac{1}{H(o)}+\sum\frac{e^{p_jt}}{p_jH'(p_j)} \tag{54}$$

which is the Expansion Theorem Solution.

We shall not attempt to discuss here cases where the expansion solution breaks down though such cases exist. In every such case, however, the breakdown is due to the failure of the impedance function $H(p)$ to satisfy the conditions necessary for the partial fraction expansion (40), and correlatively the non-existence of a solution in normal vibrations. Furthermore, it is usually possible by simple modification to deduce a modified expansion solution. It may be added here, that while the proof given above is also limited implicitly to finite networks, the expansion solution is valid in most transmission line problems.

Let us now illustrate how the expansion solution works by applying it to a few simple examples. Take first the case considered in the preceding chapter in connection with the power series solution. Required the charge Q on a condenser C in series with an inductance L and resistance R in response to a "unit e.m.f." The operational equation is

$$Q=\frac{1}{Lp^2+Rp+1/C}$$

or

$$Q=\frac{1}{L}\cdot\frac{1}{p^2+2\alpha p+\omega^2}$$

where $\alpha = R/2L$ and $\omega^2 = 1/LC$.

The roots of the equation $H(p) = o$ are the roots of the equation

$$p^2 + 2\alpha p + \omega^2 = o$$

whence

$$p_1 = -\alpha + \sqrt{\alpha^2 - \omega^2} = -\alpha + \beta,$$

$$p_2 = -\alpha - \sqrt{\alpha^2 - \omega^2} = -\alpha - \beta.$$

Also $H'(p) = 2L(p + \alpha)$, so that

$$H'(p_1) = 2\beta L$$

$$H'(p_2) = -2\beta L$$

and

$$1/H(o) = 1/L\omega^2 = C.$$

Inserting these expressions in the Expansion Theorem Solution (38), we get

$$Q = C - \frac{e^{-\alpha t}}{2\beta L}\left(\frac{e^{\beta t}}{\alpha - \beta} - \frac{e^{-\beta t}}{\alpha + \beta}\right).$$

It is now easy to verify the fact that this solution satisfies the differential equations and the boundary condition $Q = o$ and $dQ/dt = o$ at time $t = o$.

If $\omega > \alpha$, β is a pure imaginary

$$\beta = i\omega\sqrt{1 - (\alpha/\omega)^2} = i\omega'$$

and

$$Q = C - \frac{e^{-\alpha t}}{\omega' L} \cdot \frac{\omega' \cos \omega' t + \alpha \sin \omega' t}{\alpha^2 + \omega'^2}.$$

In connection with this problem we note two advantages of the expansion solution, as compared with the power series solution: (1) it is much simpler to derive from the operational equation, and (2) its numerical computation is enormously easier. A table of exponential and trigometric functions enables us to evaluate Q for any value of t almost at once whereas in the case of the power series solution the labor of computation for large values of t is very great. A third and very important advantage of the expansion solution in this particular problem is that without detailed computation we can deduce by mere inspection the general character of the function and the effect of the circuit parameters on its form: an advantage which never attaches to the power series solution.

This last property of the particular solution above is extremely important. The ideal form of solution, particularly in technical problems, is one which permits us to infer the general character and properties of the function and the effect of the circuit constants on its form, without detailed solutions. A solution which possesses these properties, even if its exact computation is not possible without prohibitive labor, is far superior to a solution which, while completely computable, tells us nothing without detailed computation. It is for this reason that some of the derived forms of solution, discussed later, are of such importance. In fact a solution which requires detailed computation before it yields the information implied in it is merely equivalent to an experimentally determined solution.

Unfortunately the advantages attaching to the expansion solution of the specific problem just discussed, do not, in general, characterize the expansion solution. The following disadvantages should be noted. First, the location of the roots of the impedance function $H(p)$ is practically impossible in the case of arbitrary networks of more than a few degrees of freedom. In the second place, when the number of degrees of freedom is large it is not only impossible to deduce the significance of the solution by inspection, but the computation becomes extremely laborious. In such cases, the practical value of the expansion solution depends, just as in the power series solution, on the possibility of recognizing and summing the expansion. This will be clear in the case of transmission lines, where the roots of $H(p)$ are infinite in number and the direct computation of the expansion solution (except in the case of the non-inductive cable) is quite impossible.

CHAPTER IV

Some General Formulas and Theorems for the Solution of Operational Equations

We have seen that the operational equation

$$h = 1/H(p)$$

is the symbolic or short-hand equivalent of the integral equation

$$\frac{1}{pH(p)} = \int_0^\infty h(t)e^{-pt}dt$$

and from the latter we have deduced two very important forms of the Heaviside solution. In recognizing the equivalence of these two equations we have a very great advantage and are able, in fact, to base the Operational Calculus on deductive instead of inductive reasoning. In this chapter we shall employ this equivalence to establish certain general formulas and theorems for the solution of operational equations. That is to say, we shall make use of the principles that (1) any method applicable to the solution of the integral equation supplies us with a corresponding method for the solution of the operational equation, and (2) a solution of any specific integral equation gives at once the solution of the corresponding operational equation. We turn therefore to a brief discussion of the appropriate methods for solving the integral equation.

It may be said at the outset, that the solution of the integral equation, like the evaluation of integrals, is a matter of considerable art and experience; in other words there is not, in general, a straightforward procedure corresponding to the process of differentiation.

On the other hand, as a purely mathematical question, it is always possible to invert the integral equation and write down $h(t)$ as an explicit function in the form of an infinite integral. For example it may be shown from the Fourier Integral that

$$h(t) = \frac{2}{\pi}\int_0^\infty \frac{\alpha(\omega)}{\omega} \sin t\omega . d\omega$$

where $\alpha(\omega)$ is defined by

$$\frac{1}{H(i\omega)} = \alpha(\omega) + i\,\beta(\omega).$$

Later on we shall briefly consider the Fourier Integral; for the present the preceding formula will not be considered further. In certain problems it is of value; for the explicit derivation of $h(t)$, however, it is usually too complicated to be of any use except in the hands of professional mathematicians. As a matter of fact, a direct attack on this formula would be equivalent to abandoning the unique simplicity and advantages of the whole Operational Calculus.

It has been noted above that any solution of the integral equation supplies a solution of the corresponding operational equation. This principle enables us to take advantage of the fact that a very large number of infinite integrals of the type

$$\int_0^\infty f(t) e^{-pt} dt$$

have been evaluated. *The evaluation of every infinite integral of this type supplies us, therefore, with the solution of an operational equation.*

Of course, not all the operational equations so solvable have physical significance. Many, however, do. Below is a list of infinite integrals with their known solutions, accompanied by the corresponding operational equation and its explicit solution. All of these solutions are directly applicable to important technical problems. It may be remarked in passing that the infinite integrals have for the most part been evaluated by advanced mathematical methods which need not concern us here.

Table of Infinite Integrals, the Corresponding Operational Equations, and Their Explicit Solutions

(a) $$\int_0^\infty e^{-pt}e^{-\lambda t}dt = \frac{1}{p+\lambda},$$

$$h = \frac{p}{p+\lambda} = e^{-\lambda t}.$$

(b) $$\int_0^\infty e^{-pt}\frac{t^n}{n!}dt = 1/p^{n+1},$$

$$h = \frac{1}{p^n} = t^n/n!.$$

(c) $$\int_0^\infty e^{-pt}\frac{1}{\sqrt{\pi t}}dt = \frac{1}{\sqrt{p}},$$

$$h = \sqrt{p} = 1/\sqrt{\pi t}.$$

(d) $$\int_0^\infty e^{-pt}\frac{(2t)^n}{1.3.5\ldots(2n-1)}\frac{dt}{\sqrt{\pi t}} = \frac{1}{p^n\sqrt{p}},$$

$$h = \frac{\sqrt{p}}{p^n} = \frac{(2t)^n}{1.3.5\ldots(2n-1)}\frac{1}{\sqrt{\pi t}}.$$

(e) $$\int_0^\infty e^{-pt}\frac{t^n}{n!}e^{-\lambda t}dt = \frac{1}{(p+\lambda)^{n+1}},$$

$$h = \frac{p}{(p+\lambda)^{n+1}} = \frac{t^n}{n!}e^{-\lambda t}.$$

(f) $$\int_0^\infty e^{-pt}\sqrt{\frac{\lambda}{\pi}}\frac{e^{-\lambda/t}}{t\sqrt{t}}dt = e^{-2\sqrt{\lambda p}},$$

$$h = pe^{-2\sqrt{\lambda p}} = \sqrt{\frac{\lambda}{\pi}}\frac{e^{-\lambda/t}}{t\sqrt{t}}.$$

(g) $$\int_0^\infty e^{-pt}\frac{e^{-\lambda/t}}{\sqrt{\pi t}}dt = \frac{e^{-2\sqrt{\lambda p}}}{\sqrt{p}},$$

$$h = \sqrt{p}\,e^{-2\sqrt{\lambda p}} = \frac{e^{-\lambda/t}}{\sqrt{\pi t}}.$$

(h) $$\int_0^\infty e^{-pt} \sin \lambda t \, dt = \frac{\lambda}{p^2+\lambda^2},$$

$$h = \frac{p\lambda}{p^2+\lambda^2} = \sin \lambda t.$$

(i) $$\int_0^\infty e^{-pt} \cos \lambda t \, dt = \frac{p}{p^2+\lambda^2},$$

$$h = \frac{p^2}{p^2+\lambda^2} = \cos \lambda t.$$

(j) $$\int_0^\infty e^{-pt}e^{-\mu t} \cos \lambda t \, dt = \frac{p+\mu}{(p+\mu)^2+\lambda^2},$$

$$h = \frac{p^2+\mu p}{(p+\mu)^2+\lambda^2} = e^{-\mu t} \cos \lambda t.$$

(k) $$\int_0^\infty e^{-pt}e^{-\mu t} \sin \lambda t \, dt = \frac{\lambda}{(p+\mu)^2+\lambda^2},$$

$$h = \frac{p\lambda}{(p+\mu)^2+\lambda^2} = e^{-\mu t} \sin \lambda t.$$

(l) $$\int_0^\infty e^{-pt} J_o(\lambda t) dt = \frac{1}{\sqrt{p^2+\lambda^2}},$$

$$h = \frac{p}{\sqrt{p^2+\lambda^2}} = J_o(\lambda t).$$

(m) $$\int_\lambda^\infty e^{-pt} J_o(\sqrt{t^2-\lambda^2}) dt = \frac{e^{-\lambda\sqrt{p^2+1}}}{\sqrt{p^2+1}},$$

$$h = \frac{p}{\sqrt{p^2+1}} e^{-\lambda\sqrt{p^2+1}} = o \text{ for } t<\lambda$$

$$= J_o(\sqrt{t^2-\lambda^2}) \text{ for } t \geq \lambda.$$

(n) $$\int_0^\infty e^{-pt} J_n(\lambda t) dt = \frac{1}{r}\left(\frac{r-p}{\lambda}\right)^n, \quad r^2 = p^2+\lambda^2,$$

$$h = \frac{p}{r}\left(\frac{r-p}{\lambda}\right)^n = J_n(\lambda t).$$

(p) $$\int_0^\infty e^{-pt}e^{\lambda t} I_o(\lambda t) dt = \frac{1}{\sqrt{p^2+2\lambda p}},$$

$$h = \frac{1}{\sqrt{1+2\lambda/p}} = e^{-\lambda t} I_o(\lambda t).$$

In formulas (l), (m), (n), $J_n(x)$ denotes the Bessel function of order n and argument x. In formula (p), $I_0(x)$ denotes the Bessel function $J_0(ix)$ where $i=\sqrt{-1}$.

This list might be greatly extended. As it is, we are in possession of a set of solutions of operational equations which occur in important technical problems and which will be employed later.

The foregoing emphasize the practical and theoretical importance of recognizing the equivalence of the integral and operational equations. With this equivalence in mind, the solution of an operational equation is often reduced to a mere reference to a table of infinite integrals. Heaviside did not recognize this equivalence. As a consequence many of his solutions of transmission line problems are extremely laborious and involved and in the end unsatisfactory because expressed in involved power series.

Not all the infinite integrals corresponding to the operational equations of physical problems have been evaluated or can be recognized without transformation. This statement corresponds exactly with the fact that a table of integrals is not always sufficient but must be supplemented by general methods of integration. We turn, therefore, to stating and discussing some general Theorems applicable to the solution of Operational Equations.

In the derivation of the operational theorems, which constitute the general rules of the Operational Calculus, the following proposition, due to Borel and known as Borel's theorem, will be frequently employed.*

If the functions $f(t)$, $f_1(t)$, and $f_2(t)$ are defined by the integral equations

$$F(p)=\int_0 f(t)e^{-pt}dt$$

$$F_1(p)=\int_0^\infty f_1(t)e^{-pt}dt$$

$$F_2(p)=\int_0^\infty f_2(t)e^{-pt}dt$$

and if the functions F, F_1 and F_2 satisfy the relation

$$F(p)=F_1(p).F_2(p)$$

* For a proof of this important theorem the reader is referred to Borel, "Lecons sur les Séries Divergentes" (1901), p. 104; to Bromwich, "Theory of Infinite Series," pp. 280-281; or to Ford, "Studies on Divergent Series and Summability," pp. 93-94 (being Vol. II of the Michigan University Science Series, published by Macmillan). The proof depends on Jacobi's transformation of a double integral: see Edward's "Integral Calculus," 1922, Vol. II, pp. 14-15.

then

$$f(t) = \int_0^t f_1(\tau) f_2(t-\tau) d\tau$$

$$= \int_0^t f_2(\tau) f_1(t-\tau) d\tau.$$

The operational theorems will now be stated and briefly proved from the integral equation identity.

Theorem I

If in the Operational Equation

$$h = 1/H(p)$$

the generalized impedance function $H(p)$ can be expanded in a sum of terms, thus

$$\frac{1}{H(p)} = \frac{1}{H_1(p)} + \frac{1}{H_2(p)} + \ldots + \frac{1}{H_n(p)},$$

and if the auxiliary operational equations

$$h_1 = \frac{1}{H_1(p)}$$

$$h_2 = \frac{1}{H_2(p)}$$

- - - - - - -

can be solved, then

$$h = h_1 + h_2 + \ldots + h_n.$$

This theorem is too obvious to require detailed proof: in fact it is self evident. The power series and expansion theorem solutions are examples of its application. In general, however, the appropriate form of expansion of $1/H(p)$ will depend on the particular problem in hand. The theorem, as it stands is a formal statement of the fact that solutions can often be obtained by an appropriate expansion whereas the equation cannot be solved as it stands.

Theorem II

If $h = h(t)$ and $g = g(t)$ are defined by the operational equations

$$h = 1/H(p)$$

$$g = 1/pH(p)$$

then

$$g(t) = \int_0^t h(\tau) d\tau.$$

To prove this theorem we start with the integral equations

$$\frac{1}{pH(p)} = \int_0^\infty h(t)e^{-pt}dt,$$

$$\frac{1}{p^2H(p)} = \int_0^\infty g(t)e^{-pt}dt.$$

The second of these is in form for an immediate application of Borel's theorem since

$$\frac{1}{p^2H(p)} = \frac{1}{p} \cdot \frac{1}{pH(p)}.$$

The functions f_1 and f_2 of Borel's theorem then satisfy the equations

$$\frac{1}{p} = \int_0^\infty f_1(t)e^{-pt}dt,$$

$$\frac{1}{pH(p)} = \int_0^\infty f_2(t)e^{-pt}dt.$$

It follows at once that

$$f_1(t) = 1$$

$$f_2(t) = h(t)$$

whence by Borel's theorem

$$g(t) = \int_0^t h(\tau)d\tau.$$

Theorem III

If $h = h(t)$ and $g = g(t)$ are defined by the operational equations

$$h = 1/H(p)$$

$$g = p/H(p)$$

then

$$g(t) = \frac{d}{dt}h(t)$$

provided $h(o) = o$.

The integral equations of the problem are

$$\frac{1}{pH(p)} = \int_0^\infty h(t)e^{-pt}dt,$$

$$\frac{1}{H(p)} = \int_0^\infty g(t)e^{-pt}dt.$$

Integrating the first of these by parts we have,

$$\frac{1}{pH(p)}=\frac{1}{p}h(o)+\frac{1}{p}\int_0^\infty h'(t)e^{-pt}dt$$

where $h'(t)=d/dt\,h(t)$.

If $h(o)=o$, we have at once

$$\frac{1}{H(p)}=\int_0^\infty h'(t)e^{-pt}dt.$$

Comparison with the integral equation for $g(t)$ shows at once that $g(t)=h'(t)$, since the integral equation determines the function uniquely.

Theorems II and III establish the characteristic Heaviside Operations of replacing $1/p$ by $\int_0^t dt$ and p by d/dt.

Theorem IV

If in the operational equation

$$h=1/H(p)$$

the generalized impedance function can be factored in the form

$$H(p)=H_1(p)\cdot H_2(p)$$

and if the auxiliary operational equations

$$h_1=1/H_1(p)$$

$$h_2=1/H_2(p)$$

define the auxiliary variables h_1 and h_2, then

$$h(t)=\frac{d}{dt}\int_0^t h_1(\tau)h_2(t-\tau)d\tau$$

$$=\frac{d}{dt}\int_0^t h_2(\tau)h_1(t-\tau)d\tau.$$

This theorem is immediately deducible from Borel's theorem and theorems II and III, as follows.

The integral equations are

$$\frac{1}{pH(p)}=p\frac{1}{pH_1(p)}\cdot\frac{1}{pH_2(p)}=\int_0^\infty h(t)e^{-pt}dt$$

$$\frac{1}{pH_1(p)}=\int_0^\infty h_1(t)e^{-pt}dt$$

$$\frac{1}{pH_2(p)}=\int_0^\infty h_2(t)e^{-pt}dt.$$

Now define an auxiliary function $g(t)$ by the operational equation

$$g=\frac{1}{pH(p)}.$$

Then

$$\frac{1}{pH_1(p)}\cdot\frac{1}{pH_2(p)}=\int_0^\infty g(t)e^{-pt}dt$$

and by Borel's theorem

$$g(t)=\int_0^t h_1(\tau)h_2(t-\tau)d\tau$$

$$=\int_0^t h_2(\tau)h_1(t-\tau)d\tau.$$

From this equation it follows that $g(o)=o$, and hence comparing the operational equations for h and g, we have by aid of Theorem III

$$h(t)=\frac{d}{dt}g(t)$$

and hence

$$h(t)=\frac{d}{dt}\int_0^t h_1(\tau)h_2(t-\tau)d\tau$$

$$=\frac{d}{dt}\int_0^t h_2(\tau)h_1(t-\tau)d\tau.$$

This theorem is extremely important, although not stated or employed by Heaviside himself. We shall make use of it in establishing two important general theorems and shall have frequent occasion to employ it in specific problems occurring in connection with the subsequent discussion of transmission theory.

Theorem V

If $h=h(t)$ and $g=g(t)$ are defined by the operational equations

$$h=\frac{1}{H(p)}$$

$$g=\frac{1}{H(p+\lambda)}$$

where λ is a positive real parameter, then

$$g(t)=\left(1+\lambda\int_0^t dt\right)e^{-\lambda t}h(t).$$

To prove this theorem we start with the integral equations

$$\frac{1}{pH(p)}=\int_0^\infty h(t)e^{-pt}dt$$

$$\frac{1}{pH(p+\lambda)}=\int_0^\infty g(t)e^{-pt}dt.$$

In the first of these equations replace the symbol p by $q+\lambda$: we get

$$\frac{1}{q+\lambda}\cdot\frac{1}{H(q+\lambda)}=\int_0^\infty h(t)e^{-\lambda t}e^{-qt}dt$$

and then to preserve our original notation replace the symbol q by p, whence

$$\frac{1}{(p+\lambda)H(p+\lambda)}=\int_0^\infty h(t)e^{-\lambda t}e^{-pt}dt. \qquad \text{(a)}$$

The integral equation in $g(t)$ can be written as

$$\left(1+\frac{\lambda}{p}\right)\frac{1}{(p+\lambda)H(p+\lambda)}=\int_0^\infty g(t)e^{-pt}dt. \qquad \text{(b)}$$

Comparing equations (a) and (b) it follows at once from theorems I and II that

$$g(t)=\left(1+\lambda\int_0^t dt\right)h(t)e^{-\lambda t}.$$

From the foregoing, the following auxiliary theorem is immediately deducible.

Theorem Va

If $h=h(t)$ and $g=g(t)$ are defined by the operational equations

$$h=\frac{1}{H(p)}$$

$$g=\frac{p}{(p+\lambda)H(p+\lambda)}$$

then

$$g(t)=h(t)e^{-\lambda t}.$$

The proof of this theorem will be left as an exercise to the reader.

Theorem VI

If $h=h(t)$ and $g=g(t)$ are defined by the operational equations

$$h=1/H(p)$$

$$g=1/H(\lambda p)$$

where λ is a positive real parameter, then

$$g(t)=h(t/\lambda).$$

We start with the integral equations

$$\frac{1}{pH(p)}=\int_0^\infty h(t)e^{-pt}dt$$

$$\frac{1}{pH(\lambda p)}=\int_0^\infty g(t)e^{-pt}dt$$

and in the first of these equations we replace p by λq and t by τ/λ, whence it becomes

$$\frac{1}{qH(\lambda q)}=\int_0^\infty h\left(\frac{\tau}{\lambda}\right)e^{-q\tau}d\tau.$$

Now replacing the symbols q and τ by p and t respectively, we have

$$\frac{1}{pH(\lambda p)}=\int_0^\infty h(t/\lambda)e^{-pt}dt$$

whence by comparison with the integral equation in $g(t)$ it follows at once that

$$g(t)=h(t/\lambda).$$

This theorem is often useful in making a convenient change in the time scale and eliminating superfluous constants.

Theorem VII

If $h=h(t)$ and $g=g(t)$ are defined by the operational equations

$$h=\frac{1}{H(p)}$$

$$g=\frac{e^{-\lambda p}}{H(p)}$$

where λ is a positive real quantity, then

$$g(t)=o \text{ for } t<\lambda$$

$$=h(t-\lambda) \text{ for } t\geq\lambda.$$

This is a very important theorem in connection with transmission line problems where retardation, due to finite velocity of propagation, occurs. Its proof proceeds as follows:

If the auxiliary function $k = k(t)$ is defined by the operational equation

$$k = e^{-\lambda p}$$

then by Theorem IV,

$$g(t) = \frac{d}{dt}\int_0^t k(\tau)h(t-\tau)\,d\tau. \tag{a}$$

Now, corresponding to the operational equation $k = e^{-\lambda p}$ we have the integral equation

$$\frac{e^{-\lambda p}}{p} = \int_0^\infty k(t)e^{-pt}dt.$$

The solution of this integral equation, which is easily verified by direct substitution in the infinite integral, is

$$\begin{aligned} k(t) &= o \text{ for } t<\lambda \\ &= 1 \text{ for } t \geq \lambda. \end{aligned}$$

Hence equation (a) becomes

$$\begin{aligned} g(t) &= o \text{ for } t<\lambda \\ &= \frac{d}{dt}\int_\lambda^t h(t-\tau)d\tau \text{ for } t\geq\lambda \\ &= h(t-\lambda) \text{ for } t\geq\lambda. \end{aligned}$$

Theorem IV, employed in the preceding proof, as stated above, is extremely important and we shall have frequent occasion to employ it in specific problems. We shall now apply it to deduce an important theorem which extends the operational calculus to arbitrary impressed forces, whereas heretofore the operational equation $h = 1/H(p)$ applied only to the case of a "unit e.m.f." impressed on the system.

It will be recalled from a previous chapter that if $x(t)$ denotes the response of a network to an arbitrary force $f(t)$, impressed at time $t=o$, and if $h(t)$ denotes the corresponding response to a "unit e.m.f.," then

$$x(t) = \frac{d}{dt}\int_0^t h(\tau)f(t-\tau)d\tau \tag{31}$$

and

$$\frac{1}{pH(p)} = \int_0^\infty h(t)e^{-pt}dt. \tag{30}$$

Now $f(t)$ may be of such form that the infinite integral

$$\int_0^\infty f(t)e^{-pt}dt$$

can be evaluated and has the value $F(p)/p$: thus

$$\int_0^\infty f(t)e^{-pt}dt = \frac{1}{p}F(p). \tag{55}$$

This is possible, of course, for many important types of applied forces, including the sinusoidal.

It follows at once from Theorem IV that $x(t)$ satisfies and is determined by the integral equation

$$\frac{1}{p}\frac{F(p)}{H(p)} = \int_0^\infty x(t)e^{-pt}dt. \tag{56}$$

We have thus succeeded, by virtue of Theorem IV in expressing the response of a network to an arbitrary e.m.f. impressed at time $t=o$, by an integral equation of the same form as that expressing the response to a "unit e.m.f." That is to say we have, at least formally, extended the operational calculus explicitly to the case of arbitrary impressed forces.

We now translate the foregoing into the corresponding Operational Theorem.

Theorem VIII

If the operational equation

$$h = 1/H(p)$$

expresses the response of a network to a "unit e.m.f." and if an arbitrary e.m.f. E impressed at time $t=o$, is expressible by the operational equation

$$E = V(p)$$

or the infinite integral

$$\int_0^\infty E(t)e^{-pt}dt = \frac{V(p)}{p}$$

then the response x of the network to the arbitrary force is given by the operational equation

$$x = \frac{V(p)}{H(p)},$$

and $x(t)$ is determined by the integral equation

$$\frac{1}{p}\frac{V(p)}{H(p)} = \int_0^\infty x(t)e^{-pt}dt.$$

Theorem IX

If the operational equation

$$h = 1/H(p)$$

is reducible to the form

$$h = \frac{F(p)}{1+\lambda K(p)}$$

where λ *is a real parameter, and if the auxiliary functions* $f=f(t)$ *and* $k=k(t)$ *are defined by the auxiliary operational equations*

$$f = F(p)$$

$$k = K(p)$$

then $h(t)$ *is determined by the Poissan Integral equation*

$$h(t) = f(t) - \lambda \int_0^t h(\tau)k(t-\tau)d\tau.$$

This theorem is of considerable practical importance in connection with the approximate and numerical solution of operational equations when the operational equation and the equivalent Laplace integral equation prove refractory. In such cases, as will be shown later, the numerical solution of the Poissan integral equations can often be rapidly and accurately effected, and in many cases the qualitative properties of $h(t)$ can be deduced from it without detailed numerical solution.

The proof of this theorem proceeds as follows:

By virtue of the relation $h = 1/H(p)$ the operational equation

$$h = \frac{F(p)}{1+\lambda K(p)}$$

can be written as

$$h + \lambda \frac{K(p)}{H(p)} = F(p)$$

$$h = F(p) - \lambda \frac{K(p)}{H(p)}.$$

A direct application of Borel's theorem or Theorem IV gives at once the explicit equivalent

$$h(t) = f(t) - \lambda \int_0^t h(\tau)k(t-\tau)d\tau.$$

The preceding theorems, together with the power series and expansion theorem solutions formulate the most important rules of the operational calculus, and are constantly employed in the solution of the electrotechnical problems. On the other hand, the table of infinite integrals furnishes the solution of a set of operational equations, which are of the greatest usefulness in the systematic study of propagation phenomena in transmission systems which will engage our attention. Before taking up this study, however, we shall first solve a few specific problems which will serve as an introduction to asymptotic and divergent solutions involving Heaviside's so-called "fractional differentiation."

Problem A: Current Entering the Non-Inductive Cable

The non-inductive cable is a smooth line with distributed resistance R and capacity C per unit length; for the present we neglect inductance and leakage. A consideration of cable problems leads to some of the most interesting questions relating to operational methods, particularly to questions regarding divergent expansions. It would seem best to allow specific problems to serve as an introduction to these general questions.

The differential equations of the cable are

$$RI = -\frac{\partial}{\partial x}V$$
$$C\frac{d}{dt}V = -\frac{\partial}{\partial x}I \tag{57}$$

where x is the distance, measured along the cable from any fixed point, I is the current at point x, and V the corresponding potential.

Replacing d/dt by the operator p, we have

$$RI = -\frac{\partial}{\partial x}V$$
$$pCV = -\frac{\partial}{\partial x}I. \tag{58}$$

Eliminating, successively, V and I from these equations, we get

$$pRCI = \frac{\partial^2}{\partial x^2}I$$

and

$$pRCV = \frac{\partial^2}{\partial x^2}V.$$

These equations have the general solutions

$$V = V_1 e^{-\gamma x} + V_2 e^{\gamma x} \quad (59)$$

$$I = \sqrt{\frac{pC}{R}}\,[V_1 e^{-\gamma x} - V_2 e^{\gamma x}] \quad (60)$$

where

$$\gamma = \sqrt{pRC}. \quad (61)$$

The term in $e^{-\gamma x}$ represents the direct wave and the term in $e^{\gamma x}$ the reflected wave. V_1 and V_2 are constants which must be so chosen as to satisfy the imposed boundary conditions at the terminals of the cable.

For the present we shall assume that the line is infinitely long so that the reflected wave is absent. We shall also assume that a voltage E is impressed directly on the cable at $x=o$: we have then,

$$V = Ee^{-x\sqrt{pCR}} = Ee^{-\sqrt{\alpha p}} \quad (62)$$

$$I = \sqrt{\frac{pC}{R}}\,Ee^{-x\sqrt{pCR}} = \sqrt{\frac{pC}{R}}\,Ee^{-\sqrt{\alpha p}} \quad (63)$$

where α denotes x^2RC.

To convert these to operational equations let us suppose that E is a "unit e.m.f." (zero before, unity after time $t=o$). We have then, in operational notation

$$V = e^{-\sqrt{\alpha p}} \quad (64)$$

$$I = \sqrt{\frac{pC}{R}}\,e^{-\sqrt{\alpha p}}. \quad (65)$$

Now suppose that $x=o$ so that $\alpha=o$, in other words consider a point at the cable terminals. Then

$$V = 1$$

$$I = \sqrt{\frac{pC}{R}}. \quad (66)$$

The first of these equations means that V is simply the impressed voltage, zero before, unity after time $t=o$, as of course, it should be from physical considerations.

Corresponding to the operational equation

$$I = \sqrt{\frac{pC}{R}}. \quad (66)$$

we have the integral equation

$$\sqrt{\frac{C}{R}}\frac{1}{\sqrt{p}}=\int_0^\infty I(t)e^{-pt}dt. \tag{67}$$

The solution of this is known (see formula (c) of the preceding table of integrals): it is

$$I=\sqrt{\frac{C}{R}}\frac{1}{\sqrt{\pi t}}=\sqrt{\frac{C}{\pi Rt}}. \tag{68}$$

Heaviside arrived at this solution from considering the known solution of the same problem in the theory of heat flow. He therefore inferred that the operational equation

$$I=\sqrt{p}$$

has the explicit solution

$$I=1/\sqrt{\pi t}.$$

This is correct; we, however, have derived it directly from the integral equation of the problem and the known integral

$$\frac{1}{\sqrt{p}}=\int_0^\infty e^{-pt}\frac{dt}{\sqrt{\pi t}}. \tag{69}$$

We then see from the foregoing that, if a "unit e.m.f." is impressed on the cable terminals, the current entering the cable is initially infinite and dies away in accordance with the formula $\sqrt{C/\pi Rt}$. The case is, of course, idealized and the infinite initial value of the current results from our ignoring the distributed inductance of the cable, which, no matter how small, keeps the initial current finite, as we shall see later.

Now let us go a step farther; suppose that in addition to distributed resistance R and capacity C, the cable also has distributed leakage G per unit length. The differential equations are now

$$\begin{aligned} RI &= -\frac{\partial}{\partial x}V \\ (Cp+G)V &= -\frac{\partial}{\partial x}I. \end{aligned} \tag{70}$$

Consequently it follows that in the operational equation for the current entering the cable we need only replace Cp by $Cp+G$. Therefore, when leakage is included, equation (66) is to be replaced by

$$I=\sqrt{\frac{pC+G}{R}}=\sqrt{\frac{C}{R}}\sqrt{p+\lambda} \tag{71}$$

where $\lambda=G/C$.

The corresponding integral equation is, of course,

$$\sqrt{\frac{C}{R}}\sqrt{\frac{p+\lambda}{p}} = \int_0^\infty I(t)e^{-pt}dt. \tag{72}$$

We shall give two solutions of this problem; first the solution of the integral equation, and second the typical Heaviside solution directly from the operational equation.

Equation (72) may be written as

$$\sqrt{\frac{C}{R}}\frac{(1+\lambda/p)}{\sqrt{p+\lambda}} = \int_0^\infty I(t)e^{-pt}dt. \tag{73}$$

Now suppose that $J(t)$ is the solution of the equation

$$\frac{1}{\sqrt{p+\lambda}} = \int_0^\infty J(t)e^{-pt}dt \tag{74}$$

it follows at once from Theorems (I) and (II) of the preceding chapter that

$$I(t)=\sqrt{\frac{C}{R}}\left(1+\lambda\int_0^t dt\right)J(t). \tag{75}$$

Also from formula (c) of the table of integrals and Theorem (Va) the solution of (74) is

$$J(t)=\frac{e^{-\lambda t}}{\sqrt{\pi t}} \tag{76}$$

whence

$$I(t)=\sqrt{\frac{C}{\pi R}}\left\{\frac{e^{-\lambda t}}{\sqrt{t}}+\lambda\int_0^t\frac{e^{-\lambda t}}{\sqrt{t}}dt\right\}. \tag{77}$$

The integral appearing in (77) can not be evaluated in finite terms; it is easily expressible as a series, however, by repeated integration by parts. Thus

$$\int_0^t\frac{e^{-\lambda t}}{\sqrt{t}}dt=2\int_0^t e^{-\lambda t}d\sqrt{t}=2\sqrt{t}\,e^{-\lambda t}+2\lambda\int_0^t e^{-\lambda t}\sqrt{t}\,dt.$$

Proceeding in this way by repeated partial integration we get for the integral term of (77)

$$2\sqrt{t}\,e^{-\lambda t}\left\{1+\frac{2\lambda t}{1.3}+\frac{(2\lambda t)^2}{1.3.5}+\dots\right\}. \tag{78}$$

The straightforward Heaviside solution is obtained by expanding the operational equation as follows:

$$I=\sqrt{\frac{C}{R}}\sqrt{p+\lambda}$$

$$=\sqrt{\frac{C}{R}}\left(1+\frac{\lambda}{p}\right)^{1/2}\sqrt{p}$$

$$=\sqrt{\frac{C}{R}}\left[1+\frac{1}{2}\frac{\lambda}{p}-\frac{1}{2.4}\left(\frac{\lambda}{p}\right)^2+\frac{1.3}{2.4.6}\left(\frac{\lambda}{p}\right)^3-\ldots\right]\sqrt{p}.$$

Identifying $\sqrt{p}$ with $1/\sqrt{\pi t}$ (from known solutions of allied problems) and substituting for $1/p^n$ multiple integrations of the nth order we get

$$I=\sqrt{\frac{C}{\pi Rt}}\left\{1+\frac{(2\lambda t)}{2}-\frac{(2\lambda t)^2}{2.3.4}+\frac{1.3(2\lambda t)^3}{2.3.4.5.6}-\ldots\right\}. \tag{79}$$

It can be verified that this solution is convergent and equivalent to (77).

This problem, while simple and of minor technical interest, will serve to introduce us to the very important and interesting question of asymptotic series solutions.

An asymptotic series, for our purposes, may be defined as a series expansion of a function, which, while divergent, may be used for numerical computation, and which exhibits the behavior of the function for sufficiently large values of the argument.

Let us return to equation (77). We observe that the series solution (78) of the definite integral becomes increasingly laborious to compute as the value of t increases. This remark applies with even greater force to the Heaviside solution (79) on account of the alternating character of the series. Right here we have an excellent example of what I regard as Heaviside's exaggerated sense of the importance of series solutions as compared with definite integrals. Consider the solution in the form of (77) as compared with Heaviside's series solution (79). The former is incomparably easier to interpret and to compute, either by numerical integration or by means of an integraph or planimeter. In fact the series (79) is practically unmanageable except for small values of t.

Returning to the question of an asymptotic expansion of the solution (77), we observe that the definite integral appearing in that equation can be written as,

$$\int_0^t \frac{e^{-\lambda t}}{\sqrt{t}}dt=\int_0^\infty \frac{e^{-\lambda t}}{\sqrt{t}}dt-\int_t^\infty \frac{e^{-\lambda t}}{\sqrt{t}}dt \tag{80}$$

provided λ is positive, as it is in this case. Now the value of the infinite integral is known; it is $\sqrt{\pi/\lambda}$. Consequently

$$\int_0^t \frac{e^{-\lambda t}}{\sqrt{t}}\,dt=\sqrt{\frac{\pi}{\lambda}}-\int_t^\infty \frac{e^{-\lambda t}}{\sqrt{t}}\,dt; \tag{81}$$

furthermore,

$$\int_0^\infty \frac{e^{-\lambda t}}{\sqrt{t}}\,dt=-\frac{1}{\lambda}\int_t^\infty \frac{1}{\sqrt{t}}\,de^{-\lambda t}=\frac{1}{\lambda}\frac{e^{-\lambda t}}{\sqrt{t}}-\frac{1}{2\lambda}\int_0^\infty \frac{e^{-\lambda t}}{t\sqrt{t}}\,dt.$$

Integrating again by parts we get

$$\frac{1}{\lambda}\frac{e^{-\lambda t}}{\sqrt{t}}-\frac{1}{2\lambda^2}\frac{e^{-\lambda t}}{t\sqrt{t}}+\frac{1.3}{2^2\lambda^2}\int_0^\infty \frac{e^{-\lambda t}}{t^2\sqrt{t}}\,dt.$$

Continuing this process, we get

$$\begin{aligned}\int_t^\infty \frac{e^{-\lambda t}}{\sqrt{t}}\,dt=\frac{e^{-\lambda t}}{\lambda\sqrt{t}}\Big[1-\frac{1}{2\lambda t}+\frac{1.3}{(2\lambda t)^2}-\frac{1.3.5}{(2\lambda t)^3}\\ +\ldots+(-1)^n\frac{1.3.5\ldots(2n-1)}{(2\lambda t)^n}\Big]\end{aligned} \tag{82}$$

$$-\frac{(-1)^n}{\lambda}\frac{1.3.5\ldots(2n+1)}{2(2\lambda)^n}\int_t^\infty \frac{e^{-\lambda t}}{t^{n+1}\sqrt{t}}\,dt.$$

Now this series is divergent, that is, if we continue out far enough in the series the terms begin to increase in value without limit. On the other hand, if we stop with the nth term the error is represented by the integral term in (82) and this is *less than*

$$\frac{(-1)^n}{\lambda\sqrt{t}}\frac{1.3.5\ldots(2n-1)}{(2\lambda t)^{n-1}}e^{-\lambda t}. \tag{83}$$

Consequently *the error committed in stopping with any term in the series is less than the value of that term.* Therefore if we stop with the smallest term in the series, the error is less than the smallest term and decreases with increasing values of t.

We can therefore write the solution (77) as

$$I\sim\sqrt{\frac{\lambda C}{R}}+\sqrt{\frac{C}{\pi Rt}}e^{-\lambda t}\left\{\frac{1}{2\lambda t}-\frac{1.3}{(2\lambda t)^2}+\frac{1.3.5}{(2\lambda t)^3}-\ldots\right\}. \tag{84}$$

The first term, since $\lambda=G/C$, is simply $\sqrt{G/R}$, the d.c. admittance of the leaky cable. The divergent series shows how the current approaches this final steady value.

In this particular problem no asymptotic solution is derivable directly from the operational equation, at least by the straightforward Heaviside processes. Asymptotic solutions, however, constitute a large and important part of Heaviside's transmission line solutions. We shall therefore discuss next a problem for which Heaviside obtained both convergent and divergent series expansions.

Problem B: Terminal Voltage on Cable with "Unit E.M.F." Impressed on Cable Through Condenser

We now take up a problem for which Heaviside obtained a divergent solution, and which will introduce us to the theory of his divergent solutions and so-called "fractional differentiation." We suppose a "unit e.m.f." impressed on an infinitely long cable of distributed resistance R and capacity C per unit length through a condenser of capacity C_o: required the voltage V at the cable terminals. The operational equation of the problem is derived as follows:—

We know from the problem just discussed that the current entering the cable whose terminal voltage is V, is, in operational notation

$$\sqrt{\frac{Cp}{R}}V.$$

But the current flowing into the condenser is

$$C_o p(1-V)$$

since the voltage across the condenser is $1-V$. Equating these two expressions we get

$$V=\frac{pC_o}{pC_o+\sqrt{pC/R}} \tag{85}$$

which is the operational equation of the problem.

This may be written as

$$V=\frac{1}{1+\frac{1}{C_o}\sqrt{\frac{C}{R}}\frac{1}{\sqrt{p}}} \tag{85}$$

$$=\frac{1}{1+\sqrt{a/p}},$$

where

$$\sqrt{a}=\frac{1}{C_o}\sqrt{C/R}.$$

Now expanding this by the binomial theorem

$$V = 1 - \sqrt{\frac{a}{p}} + \frac{a}{p} - \frac{a}{p}\sqrt{\frac{a}{p}} + \left(\frac{a}{p}\right)^2 - \cdots$$

$$= 1 + \frac{a}{p} + \left(\frac{a}{p}\right)^2 + \ldots .$$

$$-(1 + \frac{a}{p} + \left(\frac{a}{p}\right)^2 + \ldots)\sqrt{\frac{a}{p}}.$$

$$= 1 + \frac{at}{1!} + \frac{(at)^2}{2!} + \ldots . \tag{86}$$

$$-\left(\frac{2at}{1} + \frac{(2at)^2}{1.3.} + \frac{(2at)^3}{1.3.5.} + \ldots\right)\frac{1}{\sqrt{\pi a t}}$$

by the usual Heaviside rules of "algebrizing."

It is worth while verifying this from the integral equation of the problem. We have

$$\frac{1}{p}\,\frac{1}{1+\sqrt{a/p}} = \int_0^\infty V(t)e^{-pt}dt. \tag{87}$$

The left hand side can be written as

$$\frac{1}{p-a} - \frac{1}{p-a}\sqrt{\frac{a}{p}}$$

and by the formulas and theorems given in a preceding section the solution can be recognized at once as:

$$V(t) = e^{at} - \sqrt{\frac{a}{\pi}}\,e^{at}\int_0^t \frac{e^{-a\tau}}{\sqrt{\tau}}\,d\tau. \tag{88}$$

This can also be written as

$$V(t) = \sqrt{\frac{a}{\pi}}\,e^{at}\int_t^\infty \frac{e^{-a\tau}}{\sqrt{\tau}}\,d\tau. \tag{89}$$

If the definite integral of (88) is evaluated by successive partial integrations it will be found in agreement with the Heaviside solution (86).

Now the solution (86) is in powers of t and while absolutely convergent becomes progressively more difficult to interpret and compute as the value of t increases. From (89), however, we can derive a divergent or asymptotic solution applicable both for interpretation and computation, when the value of t is sufficiently large. As in the example discussed before, the asymptotic expansion results from repeated partial integrations; thus

$$\int_t^\infty \frac{e^{-a\tau}}{\sqrt{\tau}}\,d\tau = -\frac{1}{a}\int_t^\infty \frac{1}{\sqrt{\tau}}d\,e^{-a\tau}$$

$$= \frac{e^{-at}}{a\sqrt{t}} - \frac{1}{2a}\int_t^\infty \frac{e^{-a\tau}}{\tau\sqrt{\tau}}\,d\tau$$

$$= \frac{e^{-at}}{a\sqrt{\tau}} + \frac{1}{2a^2}\int_t^\infty \frac{1}{\tau\sqrt{\tau}}\,d\,e^{-a\tau}$$

$$= \frac{e^{-at}}{a\sqrt{t}} - \frac{e^{-at}}{2a^2t\sqrt{t}} + \frac{1.3}{2^2a^2}\int_t \frac{e^{-a\tau}}{\tau^2\sqrt{\tau}}\,d\tau$$

and finally

$$\frac{e^{-at}}{a\sqrt{t}}\left\{1 - \frac{1}{2at} + \frac{1.3}{(2at)^2} - \frac{1.3.5}{(2at)^3} + \ldots\right\}. \tag{90}$$

The series (90) is divergent just as is (82) of a preceding problem and the error committed by stopping with the smallest term, is of the same character and subject to the same discussion. With this understanding we write the solution (89) as

$$V(t) \sim \frac{1}{\sqrt{\pi at}}\left\{1 - \frac{1}{2at} + \frac{1.3}{(2at)^2} - \frac{1.3.5}{(2at)^3} + \ldots\right\}. \tag{91}$$

For large values of t $(at>5)$ this series is accurately and rapidly computable. Furthermore it shows by mere inspection the behavior of $V(t)$ for large values of t, and that it ultimately approaches zero as $1/\sqrt{\pi at}$.

Let us now see how Heaviside attacked this problem and how he arrived at a divergent solution from the operational formula. Returning to the operational equation (85), it can be written as

$$V = \frac{\sqrt{p/a}}{1+\sqrt{p/a}}. \tag{92}$$

Now expand the denominator by the binomial theorem: we get formally

$$V = \left\{1 - \sqrt{\frac{p}{a}} + \frac{p}{a} - \frac{p}{a}\sqrt{\frac{p}{a}} + \left(\frac{p}{a}\right)^2 - \cdots\right\}\sqrt{\frac{p}{a}}$$

$$= \left(1 + \frac{p}{a} + \left(\frac{p}{a}\right)^2 + \ldots\right)\sqrt{\frac{p}{a}} \tag{93}$$

$$-\left(\frac{p}{a} + \left(\frac{p}{a}\right)^2 + \left(\frac{p}{a}\right)^3 + \ldots\right).$$

Heaviside's procedure at this point was as remarkable as it was successful. He first discarded the second series in integral powers of p as meaningless. He then identified $\sqrt{p}$ with $1/\sqrt{\pi t}$ and replaced p^n by d^n/dt^n in the first series, getting

$$V=\left(1+\frac{1}{a}\frac{d}{dt}+\frac{1}{a^2}\frac{d^2}{dt^2}+\ldots\right)\frac{1}{\sqrt{\pi a t}} \tag{94}$$

or, carrying out the indicated differentiation,

$$V=\frac{1}{\sqrt{\pi a t}}\left(1-\frac{1}{2at}+\frac{1.3}{(2at)^2}-\frac{1.3.5}{(2at)^3}+\ldots\right)$$

which agrees with (91).

This is a typical example of a Heaviside divergent solution for which he offered no explanation and no proof other than its practical success. His procedure in this respect is quite unsatisfactory and in particular his discarding an entire series without explanation is intellectually repugnant. We shall leave these questions for the present, however; later we shall make a systematic study of his divergent solutions and rationalize them in a satisfactory manner. First, however, we shall take up a specific problem for which Heaviside obtains a divergent solution without discarding any terms.

Problem C: Current Entering a Line of Distributed L, R and C

Consider a transmission line of distributed inductance L, resistance R, and capacity C per unit length. The differential equations of current and voltage are

$$\begin{aligned}(L\frac{d}{dt}+R)I&=-\frac{\partial}{\partial x}V\\ C\frac{d}{dt}V&=-\frac{\partial}{\partial x}I.\end{aligned} \tag{95}$$

Replacing d/dt by p, we get

$$\begin{aligned}(pL+R)I&=-\frac{\partial}{\partial x}V\\ CpV&=-\frac{\partial}{\partial x}I.\end{aligned} \tag{96}$$

Equations (96) correspond exactly with (58) for the non-inductive cable: except that we must replace R by $pL+R$. For the infinitely long line, therefore, the operational formula for the current entering the line is

$$I = \sqrt{\frac{pC}{pL+R}} V_o \tag{97}$$

where V_o is the voltage at the line terminals. If this is a "unit e.m.f." we have, as our operational equation,

$$I = \sqrt{\frac{pC}{pL+R}} \tag{98}$$

which can be written as

$$I = \sqrt{\frac{C}{L}} \frac{1}{\sqrt{1+2\lambda/p}} \tag{99}$$

where $\lambda = R/2L$.

The corresponding integral equation is

$$\sqrt{\frac{C}{L}} \frac{1}{\sqrt{p^2+2\lambda p}} = \int_0^\infty e^{-pt} I(t) dt. \tag{100}$$

From either equation (99) or (100) and formula (p) of the table of integrals, we see at once that the solution is

$$I = \sqrt{\frac{C}{L}} e^{-\lambda t} I_o(\lambda t) \tag{101}$$

where $I_o(\lambda t)$ is the Bessel function $J_o(i\lambda t)$, where $i = \sqrt{-1}$. (The function is, however, a pure real.)

Heaviside's procedure, in the absence of any correlation between the operational equation and the infinite integral, was quite different. Remarking, with reference to equation (99), that "the suggestion to employ the binomial theorem is obvious," he expands it in the form

$$I = \sqrt{\frac{C}{L}} \left\{ 1 - \frac{\lambda}{p} + \frac{1.3}{2!}\left(\frac{\lambda}{p}\right)^2 - \frac{1.3.5}{3!}\left(\frac{\lambda}{p}\right)^3 + \dots \right\} \tag{102}$$

and replaces $1/p^n$ by t^n/n in accordance with the rule discussed in preceding sections. The explicit solution is then

$$I = \sqrt{\frac{C}{L}} \left\{ 1 - \lambda t + \frac{1.3}{(2!)^2}(\lambda t)^2 - \frac{1.3.5}{(3!)^2}(\lambda t)^3 + \dots \right\} \tag{103}$$

a convergent solution in rising powers of t. As yet, however, he does not recognize this series as the power series expansion of (101), which it is. He does, however, recognize the practical impossibility of using it for computing for large values of t, and remarks "But the binomial theorem furnishes another way of expanding the operator (operational equation), viz. in rising powers of p." Thus, returning to (99), it can be written as,

$$I=\sqrt{\frac{C}{L}}\frac{\sqrt{p/2\lambda}}{\sqrt{1+p/2\lambda}}. \tag{104}$$

Now expand the denominator by the binomial theorem: we get

$$I=\sqrt{\frac{C}{L}}\left\{1-\frac{p}{4\lambda}+\frac{1.3}{2!}\left(\frac{p}{4\lambda}\right)^2-\frac{1.3.5}{3!}\left(\frac{p}{4\lambda}\right)^3+\ldots\right\}\sqrt{\frac{p}{2\lambda}}. \tag{105}$$

He now identifies $\sqrt{p/2\lambda}$ with $1/\sqrt{2\pi\lambda t}$ and replaces p^n in the series by d^n/dt^n, thus getting finally

$$I=\sqrt{\frac{C}{L}}\frac{1}{\sqrt{2\pi\lambda t}}\left\{1+\frac{1}{8\lambda t}+\frac{(1.3)^2}{2!(8\lambda t)^2}+\frac{(1.3.5)^2}{3!(8\lambda t)^3}+\ldots\right\}. \tag{106}$$

This series solution is divergent: Heaviside recognizes it, however, as the asymptotic expansion of the function $e^{-\lambda t}I_0(\lambda t)$, and thus arrives at the solution

$$I=\sqrt{\frac{C}{L}}e^{-\lambda t}I_0(\lambda t) \tag{101}$$

which we have obtained from our tables of integrals.

Now the divergent expansion (106) is the well known asymptotic expansion of the function $e^{-\lambda t}I_0(\lambda t)$, which is usually derived by difficult and intricate processes. The directness and simplicity with which Heaviside derives it is extraordinary.

We note in this example that no integral powers of p appear in the divergent expansion: consequently no terms are discarded. Otherwise Heaviside's process is as startling and remarkable as in the example discussed in the preceding section.

We shall later encounter many problems in which asymptotic solutions are derivable as in the preceding example. We have sufficient data, however, in these two typical examples to take up a systematic discussion of the theory of Heaviside's divergent solution of the operational equation.

CHAPTER V

The Theory of the Asymptotic Solution of Operational Equations

A study of Heaviside's methods, as exemplified in the preceding examples and in many problems dealt with in his Electromagnetic

Theory, Vol. II, shows that they may be divided into two classes: (I) those of which the operational equation is of the form

$$h = F(p)\sqrt{p} \tag{I}$$

and (II) those of which the operational equation is of the form

$$h = \phi(p^k\sqrt{p}) \tag{II}$$

where k is an integer.

Heaviside himself does not distinguish between the two classes, but employs the following rule for obtaining asymptotic expansion solutions:

If the operational equation

$$h = 1/H(p)$$

can be expanded in the form

$$h = a_o + a_1p + a_2p^2 + \,.\,.\, + a_np^n + \,.\,.\,.$$

$$(b_o + b_1p + b_2p^2 + \,.\,.\, + b_np^n + \,.\,.\,.\,)\sqrt{p}\,, \tag{107}$$

a solution, usually divergent, is obtained by discarding the first expansion entirely, except for the leading constant terms a_o, *replacing* $\sqrt{p}$ *by* $1/\sqrt{\pi t}$ *and* p^n *by* d^n/dt^n *in the second expansion, whence an explicit series solution results.*

$$h = a_o + \left(b_o + b_1\frac{d}{dt} + b_2\frac{d^2}{dt^2} + \,.\,.\,.\right)\frac{1}{\sqrt{\pi t}} \tag{108}$$

$$= a_o + \frac{1}{\sqrt{\pi t}}\left(b_o - b_1\frac{1}{2t} + b_2\frac{1.3}{(2t)^2} - b_3\frac{1.3.5}{(2t)^3} + \,.\,.\,.\right). \tag{109}$$

It should be expressly understood that Heaviside nowhere himself states this rule formally. He does not distinguish between the two cases where integral series in p do and do not appear, although very important mathematical distinctions are involved. Furthermore, in one case he modifies his usual procedure by adding an extra term (Elm. Th. Vol. II, pg. 42-44). It certainly represents, however, his usual procedure in a very large number of problems.

A completely satisfactory theory of the Heaviside Rule, just stated, has not yet been arrived at although we can always verify the divergent solutions in specific problems. Furthermore, it is not as yet known just how general it is, though it certainly works successfully in a large number of physical problems to which it has been applied. Finally we know nothing in general as to the asymptotic character of the resulting expansion. In some cases it leads to an expansion in which the error is less than the last term included, in others remarkably enough the expansion is everywhere convergent, while in yet others its application leads to a series which is meaningless for a certain range of values of t.

Heaviside himself gives no information which would serve us as a guide in informing us when the rule is applicable and when it is not. Consequently it becomes a matter of practical importance, not only to investigate the underlying mathematical philosophy of the rule and to establish it on the basis of orthodox mathematics, but also to develop if possible a criterion of its applicability. In this investigation we shall have recourse to the integral equation of the problem.

We shall take up first the type of problem (Class I) in which the operational equation is

$$h=\frac{1}{H(p)}=F(p)\sqrt{p} \tag{110}$$

and assume that $F(p)$ admits of the formal power series expansion

$$F(p)=b_0+b_1p+b_2p^2+b_3p^3+\ldots \tag{111}$$

The corresponding integral equation is

$$\frac{F(p)}{\sqrt{p}}=\int_0^\infty h(t)e^{-pt}dt. \tag{112}$$

We now assume the existence of an auxiliary function $k(t)$, defined and determined by the auxiliary integral equation

$$F(p)=\int_0^\infty k(t)e^{-pt}dt. \tag{113}$$

Now since

$$\frac{1}{\sqrt{p}}=\int_0^\infty e^{-pt}\frac{dt}{\sqrt{\pi t}} \tag{114}$$

it follows from (112), (113), and (114) and Borel's Theorem, or Theorem IV, that

$$h(t)=\frac{1}{\sqrt{\pi}}\int_0^t\frac{k(\tau)}{\sqrt{t-\tau}}d\tau. \tag{115}$$

Now if we differentiate (113) repeatedly with respect to p and put $p=o$, it follows from the expansion (111) that

$$b_n=(-1)^n\int_0^\infty\frac{t^n}{n!}k(t)dt. \tag{116}$$

This equation presupposes, it should be noted, the convergence of the infinite integrals for all values of n, and therefore imposes severe restrictions on $k(t)$ and hence on $F(p)$. We shall suppose that these restrictions are satisfied, and discuss them later.

Now (115) can be written as:—

$$h(t) = \frac{1}{\sqrt{\pi t}} \int_0^t d\tau . k(\tau) \, (1 - \tau/t)^{-1/2}. \tag{117}$$

It can be shown that, if $k(t)$ satisfies the restrictions underlying (116), the integral (117) has an asymptotic solution obtained as follows:—Expand the factor $(1-\tau/t)^{-1/2}$ by the binomial theorem, replace the upper limit of integration by ∞, and integrate term by term: thus

$$h(t) \sim \frac{1}{\sqrt{\pi t}} \left\{ \int_0^\infty k(t)dt + \frac{1}{2t}\int_0^\infty \frac{t}{1!}k(t)dt + \frac{1.3}{(2t)^2}\int_0^\infty \frac{t^2}{2!}k(t)dt + \ldots \right\}. \tag{118}$$

Finally from (116) we get

$$h(t) \sim \frac{1}{\sqrt{\pi t}} \left\{ b_o - b_1 \frac{1}{2t} + b_2 \frac{1.3}{(2t)^2} - b_3 \frac{1.3.5}{(2t)^3} + \ldots \right\} \tag{119}$$

which agrees exactly with the Heaviside rule for this case.

The foregoing says nothing regarding the asymptotic character of the solution. It is easy to see qualitatively, however, that (118) and therefore (119) does represent the behavior of the definite integral (117) for large values of t, provided $k(t)$ converges with sufficient rapidity.

The foregoing analysis may now be summarized in the following proposition:

If the operational equation $h = 1/H(p)$ is reducible to the form

$$h = F(p)\sqrt{p}$$

and if $F(p)$ admits of power series expansion in p: thus

$$F(p) = b_o + b_1 p + b_2 p^2 + \ldots + b_n p^n + \ldots .$$

so that, formally,

$$h = (b_o + b_1 p + b_2 p^2 + \ldots + b_n p^n + \ldots)\sqrt{p}$$

an explicit series solution, usually asymptotic, is obtained by replacing $\sqrt{p}$ by $1/\sqrt{\pi t}$ and p^n(n integral) by d^n/dt^n, whence

$$h(t) \sim \left(b_o + b_1 \frac{d}{dt} + b_2 \frac{d^2}{dt^2} + \ldots \right) \frac{1}{\sqrt{\pi t}}$$

$$\frac{1}{\sqrt{\pi t}} \left(b_o - b_1 \frac{1}{2t} + b_2 \frac{1.3}{(2t)^2} - b_3 \frac{1.3.5}{(2t)^3} + \ldots \right)$$

provided the function $k = k(t)$, defined by the operational equation $k = F(p)$, and the infinite integrals

$$\int_0^\infty t^n k(t)dt \qquad (n = 1, 2, \dots)$$

exist.

We shall now apply the foregoing theory to a physical problem discussed in the last section: namely, the current entering an infinitely long line of inductance L, resistance R and capacity C per unit length. It will be recalled (see equation (100)) that the integral equation of this problem is

$$\sqrt{\frac{C}{L}}\frac{1}{\sqrt{p^2+2\lambda p}} = \int_0^\infty e^{-pt}I(t)dt$$

where $\lambda = R/2L$, and that the solution is

$$I = \sqrt{\frac{C}{L}}\, e^{-\lambda t} I_0(\lambda t).$$

We can derive the solution in another form appropriate for our purposes by writing

$$\sqrt{\frac{C}{L}}\frac{1}{\sqrt{p}}\frac{1}{\sqrt{p+2\lambda}} = \int_0^\infty e^{-pt}I(t)\,dt$$

Now since

$$\frac{1}{\sqrt{p}} = \int_0^\infty e^{-pt}\frac{dt}{\sqrt{\pi t}}$$

and

$$\frac{1}{\sqrt{p+2\lambda}} = \int_0^\infty e^{-pt}\frac{e^{-2\lambda t}}{\sqrt{\pi t}}dt$$

it follows from Borel's theorem that

$$I = \sqrt{\frac{C}{L}}\,\frac{1}{\pi}\int_0^t \frac{e^{-2\lambda\tau}}{\sqrt{\tau}\sqrt{t-\tau}}d\tau.$$

Now subject this definite integral (omitting the factor $\sqrt{C/L}$) to the same process applied to (117): we get

$$\frac{1}{\pi\sqrt{t}}\left\{\int_0^\infty \frac{e^{-2\lambda t}}{\sqrt{t}}dt + \frac{1}{2t}\int_0^\infty \frac{\sqrt{t}}{1!}e^{-2\lambda t}dt + \frac{1.3}{(2t)^2}\int_0^\infty \frac{t\sqrt{t}}{2!}e^{-2\lambda t}dt + \dots\right\}.$$

The infinite integrals are known and have been evaluated. Substituting their values this series becomes:—

$$\frac{1}{\sqrt{2\pi\lambda t}}\left\{1+\frac{1}{8\lambda t}+\frac{1^2.3^2}{2!(8\lambda t)^2}+\frac{1^2.3^2.5^2}{3!(8\lambda t)^3}+\ldots\right\}$$

which is in fact the well known asymptotic expansion of the function $e^{-\lambda t}I_o(\lambda t)$.

A second example may be worth while. Consider the case of an e.m.f. $e^{-\lambda t}$ impressed at time $t=o$ on a cable of distributed resistance R and capacity C: required the current entering the cable. The required formula is [6]

$$I=\sqrt{\frac{C}{\pi R}}\frac{d}{dt}\int_0^t\frac{e^{-\lambda(t-\tau)}}{\sqrt{\tau}}\,d\tau$$

$$=\sqrt{\frac{C}{\pi R}}\left\{\frac{1}{\sqrt{t}}-\lambda\int_0^t\frac{e^{-\lambda\tau}}{\sqrt{t-\tau}}\,d\tau\right\} \qquad (120)$$

by obvious transformations.

Asymptotic expansion of the definite integral as in the preceding example gives the asymptotic formula

$$I=-\sqrt{\frac{C}{\pi Rt}}\left\{\frac{1}{2\lambda t}+\frac{1.3}{(2\lambda t)^2}+\frac{1.3.5}{(2\lambda t)^3}+\ldots\right\}.$$

The operational formula of the problem is

$$I=\sqrt{\frac{C}{R}}\frac{p}{p+\lambda}\sqrt{p}$$

$$=\sqrt{\frac{C}{R}}\frac{p/\lambda}{1+p/\lambda}\sqrt{p}$$

$$=\sqrt{\frac{C}{R}}\left\{\frac{p}{\lambda}-\left(\frac{p}{\lambda}\right)^2+\left(\frac{p}{\lambda}\right)^3-\ldots\right\}\sqrt{p}.$$

Applying the Heaviside Rule, we get the asymptotic expansion

$$I=-\sqrt{\frac{C}{\pi Rt}}\left\{\frac{1}{2\lambda t}+\frac{1.3}{(2\lambda t)^2}+\frac{1.3.5}{(2\lambda t)^3}+\ldots\right\}$$

which agrees with the preceding formula, derived from the definite integral.

We shall now discuss a specific problem in which the Heaviside Rule breaks down. For example let us take the preceding problem, and

[6] The derivation of the formulas in this problem is left as an exercise for the reader.

replace the applied e.m.f. $e^{-\lambda t}$ by $\sin \omega t$. The formula corresponding to (120) is now

$$I=\omega\sqrt{\frac{C}{\pi R}}\int_0^t \frac{\cos \omega\tau}{\sqrt{t-\tau}}d\tau. \tag{121}$$

If we now attempt to expand the definite integral of (121) in the same way as that of (120), we find that the process breaks down because each component of the infinite integral is now itself infinite. In fact no asymptotic solution of this problem exists.

Let us, however, start with the operational formula: since

$$\int_0^\infty e^{-pt}\sin \omega t.dt=\frac{\omega}{p^2+\omega^2}$$

it is

$$I=\sqrt{\frac{C}{R}}\frac{\omega p}{p^2+\omega^2}\sqrt{p}.$$

Now expand this in accordance with the Heaviside Rule: we get, operationally,

$$I=\sqrt{\frac{C}{R}}\left\{\left(\frac{p}{\omega}\right)-\left(\frac{p}{\omega}\right)^3+\left(\frac{p}{\omega}\right)^5-\dots\right\}\sqrt{p}$$

and explicitly

$$I=-\sqrt{\frac{C}{\pi Rt}}\left\{\frac{1}{2\omega t}-\frac{1.3.5}{(2\omega t)^3}+\dots\right\}$$

which is quite incorrect.[7] The incorrectness of the result will be evident when we remember that the final value of the current is the *steady-state* current in response to $\sin \omega t$, or

$$\sqrt{\frac{\omega C}{2R}}(\cos \omega t+\sin \omega t). \tag{122}$$

This result can be derived directly from (121) by writing it as

$$I=\omega\sqrt{\frac{C}{\pi R}}\left\{\cos \omega t\int_0^t\frac{\cos \omega t}{\sqrt{t}}dt+\sin \omega t\int_0^t\frac{\sin \omega t}{\sqrt{t}}dt\right\}. \tag{123}$$

If the time is made indefinitely great the upper limits of the integrals may be replaced by infinity. The infinite integrals are known: substitution of their known values gives (122).

This example illustrates the care which must be used in applying Heaviside's rules for obtaining divergent solutions and the importance

[7] While this series is incorrect as an asymptotic expansion of the current it has important significance, as we shall see, in connection with the building up of alternating currents.

of having a method of checking the correctness of his processes and results.

We now take up the discussion of the asymptotic expansion solutions of operational equations of the type

$$h=\phi\ (p^k\sqrt{p}) \qquad (k \text{ integral}). \tag{123}$$

In this discussion we shall, as a matter of convenience, assume that $k=o$, so that the equation reduces to the form

$$h=\phi(\sqrt{p}). \tag{123a}$$

This will involve no loss of essential generality, since the analytical theory of the two equations is precisely the same.

The Heaviside Rule for this type of operational equation may be formulated as follows:

If the operational equation $h=1/H(p)$ is reducible to the form

$$h=\phi\ (p^k\sqrt{p})$$

and if ϕ admits of power series expansion in the argument, thus

$$h=a_o+a_1p^k\sqrt{p}+a_2p^{2k+1}+a_3p^{3k+1}\sqrt{p}+\ldots$$

a series solution, usually divergent and asymptotic, is obtained by discarding integral powers of p, and writing

$$h=a_o+(a_1p^k+a_3p^{3k+1}+a_5p^{5k+2}+\ldots)\sqrt{p}.$$

The explicit series solution then results from replacing $\sqrt{p}$ by $1/\sqrt{\pi t}$, and p^n by d^n/dt^n, whence

$$h\sim a_o+\left(a_1\frac{d^k}{dt^k}+a_3\frac{d^{3k+1}}{dt^{3k+1}}+a_5\frac{d^{5k+2}}{dt^{5k+2}}+\ldots\right)\frac{1}{\sqrt{\pi t}}$$

$$\sim a_o+\frac{(-1)^k}{\sqrt{\pi t}}\left(a_1\frac{1.3\ldots(2k-1)}{(2t)^k}-a_3\frac{1.3\ldots(6k+1)}{(2t)^{3k+1}}+\ldots\right).$$

The theory of this series solution will be based on the following proposition, deducible from the identity $\int_0^\infty \frac{e^{-pt}}{\sqrt{\pi t}}dt=1/\sqrt{p}$.

If the function $F(p)$ of the integral equation

$$F(p)=\int_0^\infty f(t)e^{-pt}dt$$

approaches $1/\sqrt{p}$ as p approaches zero, then $f(t)$ ultimately behaves as $1/\sqrt{\pi t}$: that is, if $F(p)\to 1/\sqrt{p}$ as $p\to o$, then $f(t)\sim 1/\sqrt{\pi t}$ as $t\to\infty$, provided that $f(t)$ converges to zero, and contains no term or factor which is ultimately oscillatory.

To illustrate what this condition means suppose that

$$f(t)=\frac{a}{\sqrt{\pi t}}+\frac{b\cos\omega t}{\sqrt{\pi t}}$$

then

$$\int_0^\infty f(t)e^{-pt}dt\to a/\sqrt{p} \text{ as } p\to 0,$$

and the oscillatory term in $f(t)$ converges to a higher order. The presence of such oscillatory terms vitiate, therefore, the Heaviside Rule: in the following discussion we shall assume that they are absent.

We are now prepared to discuss the operational equation

$$h=\phi\,(p^k\sqrt{p})$$

and for convenience shall assume that $k=0$ so that the operational equation becomes

$$h=\phi\,(\sqrt{p})$$

of which the corresponding or equivalent integral equation is

$$\frac{1}{p}\phi(\sqrt{p})=\int_0^\infty h(t)e^{-pt}dt. \qquad (123b)$$

We assume that $\phi(\sqrt{p})$ admits of formal power series expansion in the argument: thus

$$\phi(\sqrt{p})=a_0+a_1\sqrt{p}+a_2p+a_3p\sqrt{p}+a_4p^2+\ldots$$

without, however, implying anything regarding the convergence of this expansion.

We now introduce the series of auxiliary functions, $g, g_1, g_2, g_3, \ldots$ defined by the following scheme

$$\left.\begin{aligned}
g(t)&=h(t)-a_0\\
g_1(t)&=g(t)-\frac{a_1}{\sqrt{\pi t}}\\
g_2(t)&=tg_1(t)+\frac{1}{2}\frac{a_3}{\sqrt{\pi t}}\\
g_3(t)&=t.g_2(t)-\frac{1.3}{2^2}\frac{a_5}{\sqrt{\pi t}}\\
g_4(t)&=t.g_3(t)+\frac{1.3.5}{2^3}\frac{a_7}{\sqrt{\pi t}}\\
&\text{- - - - - - - - - - - - - - -}
\end{aligned}\right\} \qquad (123c)$$

Successive substitutions in the integral equation (123b) and repeated differentiations with respect to p, lead to the set of formulas,

$$\begin{aligned} &\int_0^\infty g(t)e^{-pt}dt \sim \frac{a_1}{\sqrt{p}} \text{ as } p\to 0 \\ &\int_0^\infty t.g_1(t)e^{-pt}dt \sim \frac{a_3}{2\sqrt{p}} \text{ as } p\to 0 \\ &\int_0^\infty t.g_2(t)e^{-pt}dt \sim \frac{1.3}{2^2}\frac{a_5}{\sqrt{p}} \text{ as } p\to 0 \\ &\int_0^\infty t.g_3(t)e^{-pt}dt \sim -\frac{1.3.5}{2^3}\frac{a_7}{\sqrt{p}} \text{ as } p\to 0 \\ &\text{-------------------------} \end{aligned} \qquad (123d)$$

Now *assuming* that $h(t)$ satisfies the restrictions stated in the preceding proposition, it follows from that proposition, that

$$\begin{aligned} &g(t) \sim a_1/\sqrt{\pi t} \text{ as } t\to\infty \\ &g_1(t) \sim -\frac{a_3}{2t\sqrt{\pi t}} \text{ as } t\to\infty \\ &g_2(t) \sim \frac{1.3}{2^2 t}\frac{a_5}{\sqrt{\pi t}} \text{ as } t\to\infty \\ &g_3(t) \sim -\frac{1.3.5}{2^3 t}\frac{a_7}{\sqrt{\pi t}} \text{ as } t\to\infty \\ &\text{-------------------} \end{aligned} \qquad (123e)$$

From the set equations (123d) and (123e) it follows by successive substitutions that

$$h(t) \sim a_o + \frac{1}{\sqrt{\pi t}}\left(a_1 - a_3\frac{1}{2t} + a_5\frac{1.3}{2t^2} - a_7\frac{1.3.5}{(2t)^3} + \ldots\right)$$

which agrees with the series gotten by applying the Heaviside Rule.

The defect of this derivation, which, however, appears to be inherent, is that it requires us to know or assume at the outset that $h(t)$ satisfies the required restrictions. Consequently an automatic application of the Heaviside Rule may or may not give correct results. On the other hand if we know that an expansion solution in inverse fractional powers of t exists, the Heaviside Rule gives the series with extraordinary directness and simplicity.

The type of expansion solution just discussed will now be illustrated by some specific problems. The first problem is that of the propagated voltage in the non-inductive cable in response to a "unit e.m.f". It will be recalled that in a preceding chapter we derived the operational formula

$$V = e^{-\sqrt{\alpha p}} \tag{121}$$

where $\alpha = x^2 RC$, for the voltage at distance x from the terminal of a non-inductive cable of distributed resistance R and capacity C, in response to a "unit e.m.f." impressed at point $x=0$. Heaviside's solution of this operational equation proceeds as follows:

Expansion of the exponential function in the usual power series gives

$$V = 1 - \frac{\sqrt{\alpha p}}{1!} + \frac{\alpha p}{2!} - \frac{\alpha p \sqrt{\alpha p}}{3!} + \frac{(\alpha p)^2}{4!} - \dots$$

which may be rearranged as

$$V = 1 - \left(1 + \frac{\alpha p}{3!} + \frac{(\alpha p)^2}{5!} + \dots\right)\sqrt{\alpha p} + \left(\frac{\alpha p}{2!} + \frac{(\alpha p)^2}{4!} + \frac{(\alpha p)^3}{6!} + \dots\right) \tag{125}$$

Heaviside then discards the series in integral powers of p entirely, replaces $\sqrt{p}$ by $1/\sqrt{\pi t}$ and p^n by d^n/dt^n in the first series, and then gets

$$V = 1 - \left(1 + \frac{\alpha}{3!}\frac{d}{dt} + \frac{\alpha^2}{5!}\frac{d^2}{dt^2} + \dots\right)\sqrt{\frac{\alpha}{\pi t}}$$

$$= 1 - \sqrt{\frac{\alpha}{\pi t}}\left(1 - \frac{1}{3!}\left(\frac{\alpha}{2t}\right) + \frac{1.3}{5!}\left(\frac{\alpha}{2t}\right)^2 - \frac{1.3.5}{7!}\left(\frac{\alpha}{2t}\right)^3 + \dots\right) \tag{126}$$

or

$$V = 1 - \sqrt{\frac{\alpha}{\pi t}}\left(1 - \frac{1}{3}\left(\frac{\alpha}{4t}\right) + \frac{1}{5.2!}\left(\frac{\alpha}{4t}\right)^2 - \frac{1}{7.3!}\left(\frac{\alpha}{4t}\right)^3 + \dots\right). \tag{127}$$

This solution is correct, as will be shown subsequently.

A rather remarkable feature of this solution—a point on which Heaviside makes no comment—is that it is absolutely convergent. In other words, a process of expansion which in other problems leads to a divergent or asymptotic solution, here results in a convergent series expansion.

To verify this solution we start with the corresponding integral equation of the problem

$$\frac{1}{p}e^{-\sqrt{\alpha p}} = \int_0^\infty V(t)e^{-pt}dt. \tag{128}$$

It follows from this formula and theorem (V) that

$$V(t) = \int_0^t \phi(t)dt$$

where $\phi(t)$ is determined by the integral equation

$$e^{-\sqrt{\alpha p}} = \int_0^{\infty} \phi(t) e^{-pt} dt.$$

Now from formula (f) of the table of integrals

$$e^{-\sqrt{\alpha p}} = \frac{1}{2}\sqrt{\frac{\alpha}{\pi}} \int_0^{\infty} e^{-pt} \frac{e^{-\alpha/4t}}{t\sqrt{t}} dt$$

whence

$$\phi(t) = \frac{1}{2}\sqrt{\frac{\alpha}{\pi}} \frac{e^{-\alpha/4t}}{t\sqrt{t}}$$

and finally

$$V(t) = \frac{1}{\sqrt{\pi}} \int_0^{t'} \frac{e^{-1/\tau}}{\tau\sqrt{\tau}} d\tau, \text{ where } t' = 4t/\alpha. \tag{129}$$

To convert this to the form of (127) we write

$$V(t) = \frac{1}{\sqrt{\pi}} \int_0^{\infty} \frac{e^{-1/\tau}}{\tau\sqrt{\tau}} d\tau - \frac{1}{\sqrt{\pi}} \int_{t'}^{\infty} \frac{e^{-1/\tau}}{\tau\sqrt{\tau}} d\tau. \tag{130}$$

The value of the infinite integral is known to be unity so that

$$V = 1 - \frac{1}{\sqrt{\pi}} \int_{t'}^{\infty} \frac{e^{-1/\tau}}{\tau\sqrt{\tau}} d\tau. \tag{131}$$

Now in the integral term of (131) expand $e^{-1/\tau}$ in the usual exponential power series and then integrate term by term: the series solution (127) results. This series, while absolutely convergent, is difficult to compute for small values of t; an asymptotic expansion, which can be employed for computation for small values of t is gotten as follows:—

Write (129) as

$$V = \frac{1}{\sqrt{\pi}} \int_0^{t'} \sqrt{\tau}\, de^{-1/\tau}$$

$$= \sqrt{\frac{t'}{\pi}} e^{-1/t'} - \frac{1}{2\sqrt{\pi}} \int_0^{t'} \frac{e^{-1/\tau}}{\sqrt{\tau}} d\tau.$$

Repeated partial integrations of this type lead to the series

$$V = \sqrt{\frac{t'}{\pi}} e^{-1/t'} \left\{ 1 - \left(\frac{t'}{2}\right) + 1.3\left(\frac{t'}{2}\right)^2 - \dots \right\}. \tag{132}$$

It is interesting to note, in passing, that an asymptotic solution of this type does not appear to be directly deducible from the operational equation. We observe also that, in this problem, the series in inverse powers of t is convergent while the series in ascending powers of t is divergent: the converse is the case in the problems discussed previously.

A second specific problem may be stated as follows:

Let a "unit e.m.f." be impressed on an infinitely long non-inductive cable of distributed resistance R and capacity C per unit length through a terminal resistance R_0: required the voltage V on the cable terminals. The formulation of the operational equation of this problem is very simple. It will be recalled that the operational formula for the current entering the cable with terminal voltage V is $V\sqrt{Cp/R}$. But the current is clearly also equal to $(1-V)/R_0$: equating these expressions we get

$$\frac{1-V}{R_0}=V\sqrt{pC/R}$$

whence

$$V=\frac{1}{\sqrt{p/\lambda}+1} \tag{133}$$

where $1/\sqrt{\lambda}=R_0\sqrt{C/R}$. This is the required operational formula.

To derive the Heaviside divergent expansion, expand (133) by the binomial theorem: thus

$$\begin{aligned} V&=1-\sqrt{p/\lambda}+(p/\lambda)-(p/\lambda)^{3/2}+\ldots \\ &=1-(1+p/\lambda+(p/\lambda)^2+\ldots)\sqrt{p/\lambda} \\ &\quad+(p/\lambda+(p/\lambda)^2+(p/\lambda)^3+\ldots). \end{aligned} \tag{134}$$

Discard the second series in integral powers of p; replace $\sqrt{p}$ by $1/\sqrt{\pi t}$ and p^n by d^n/dt^n in the first series, thus getting

$$V=1-\left(1+\frac{1}{\lambda}\frac{d}{dt}+\frac{1}{\lambda^2}\frac{d^2}{dt^2}+\ldots\right)\frac{1}{\sqrt{\pi\lambda t}} \tag{135}$$

$$=1-\frac{1}{\sqrt{\pi\lambda t}}\left(1-\frac{1}{2\lambda t}+\frac{1.3}{(2\lambda t)^2}-\ldots\right) \tag{136}$$

which is the asymptotic solution of the problem.

To verify this solution we shall consider the more general operational equation

$$h=\frac{1}{p^n\sqrt{p}+1}\qquad (n \text{ integral}) \tag{137}$$

a form of equation to which a number of fairly important problems is reducible. (The parameter λ of equation (133) can be eliminated from explicit consideration by means of theorem VI.)

Multiplying numerator and denominator of equation (137) by $p^n\sqrt{p}-1$, it becomes

$$h=\frac{p^n\sqrt{p}-1}{p^{2n+1}-1}=\frac{p^n}{p^{2n+1}-1}\sqrt{p}-\frac{1}{p^{2n+1}-1} \tag{138}$$

and by direct partial fraction expansion, this is equivalent to

$$h=\frac{\sqrt{p}}{2n+1}\sum_{m=o}^{2n}\frac{p_m^{n+1}}{p-p_m}-\frac{1}{2n+1}\sum_{m=o}^{2n}\frac{p_m}{p-p_m} \tag{139}$$

where

$$p_m=e^{i\frac{2m\pi}{2n+1}} \qquad (m=0,1,2\ldots 2n).$$

Write, for convenience,

$$h=\sum_{m=o}^{2n}h_m$$

and consider the operational equation

$$h_m=\frac{1}{2n+1}\left(\frac{p_m^{n+1}}{p-p_m}\sqrt{p}-\frac{p_m}{p-p_m}\right). \tag{140}$$

By the rules of the operational calculus, fully discussed in preceding chapters, the solution of this is

$$h_m(t)=\frac{1}{2n+1}\left(\frac{p_m^{n+1}}{\sqrt{\pi}}\int_0^t\frac{e^{p_m(t-\tau)}}{\sqrt{\tau}}d\tau+1-e^{p_mt}\right). \tag{141}$$

We have now to distinguish two cases: (1) when the *real part* of p_m is positive, and (2) when the real part is negative.

Taking up case (1) first, the preceding can be written

$$h_m(t)=\frac{1}{2n+1}\left(1+e^{p_mt}\left\{\frac{p_m^{n+1}}{\sqrt{\pi}}\int_0^t\frac{e^{p_m\tau}}{\sqrt{\tau}}d\tau-1\right\}\right. \tag{142}$$

$$=\frac{1}{2n+1}\left(\begin{array}{l}1+e^{p_mt}\left\{\dfrac{p_m^{n+1}}{\sqrt{\pi}}\displaystyle\int_0^\infty\dfrac{e^{-p_m\tau}}{\sqrt{\tau}}d\tau-1\right\}\\ -\dfrac{p_m^{n+1}}{\sqrt{\pi}}e^{p_mt}\displaystyle\int_t^\infty\dfrac{e^{-p_mt}}{\sqrt{\tau}}d\tau\end{array}\right) \tag{143}$$

$$=\frac{1}{2n+1}\left(1-\frac{p_m^{n+1}}{\sqrt{\pi}}e^{p_mt}\int_t^\infty\frac{e^{-p_m\tau}}{\sqrt{\tau}}d\tau\right). \tag{144}$$

Repeated integration by parts of the definite integral leads to an asymptotic series, identical with that obtained by applying the Heaviside Rule to the operational equation (137).

If, on the other hand, the *real part* of p_m is negative, we write (144) as

$$h_m(t) = \frac{1}{2n+1}\left(1 - e^{p_m t} + \frac{p_m^{n+1}}{\sqrt{\pi}} \int_0^t \frac{e^{p_m \tau}}{\sqrt{t-\tau}} d\tau \right). \tag{145}$$

The term $e^{p_m t}$ ultimately dies away, and the definite integral can be expanded asymptotically in accordance with the theory discussed under Rule I, again leading to an asymptotic series identical with that given by direct application of the Heaviside Rule to the operational equation.

Consequently since the operational equation in h_n can be asymptotically expanded by means of the Heaviside Rule, the operational equation in $h = \sum h_m$ is similarly asymptotically expandible, and the Heaviside Rule is verified for equation (133).

We have now covered, more or less completely, the theoretical rules and principles of the operational calculus in so far as they can be formulated in general terms. We shall now apply these principles and rules to the solution of important technical problems relating to the propagation of current and voltage along lines. In doing, so, while we shall take advantage of our table of integrals with the corresponding solutions of the operational equation, we shall also sketch Heaviside's own methods of solution.

We shall close this discussion of divergent and asymptotic expansions with a general expansion solution of considerable theoretical and practical importance in the problem of the building-up of alternating currents. It will be recalled from Theorem III that the response of a network of generalized operational impedance $H(p)$ to an e.m.f. $E(t)$ impressed at time $t=o$ is given by the operational formula

$$x = \frac{V(p)}{H(p)}$$

where $E = V(p)$ is the operational equation of the applied e.m.f.: that is, analytically

$$\frac{1}{p} V(p) = \int_0^\infty E(t) e^{-pt} dt.$$

Now suppose that the impressed e.m.f. is $\sin \omega t$: then by formula (h) of the table of integrals

$$V(p) = \frac{\omega p}{p^2 + \omega^2} \tag{146}$$

and denoting x by x_s

$$x_s = \frac{\omega p}{p^2+\omega^2}\frac{1}{H(p)}. \tag{147}$$

If, on the other hand, the impressed e.m.f. is $\cos \omega t$, then by formula (i)

$$V(p) = \frac{p^2}{p^2+\omega^2} \tag{148}$$

and

$$x = x_c = \frac{p^2}{p^2+\omega^2}\frac{1}{H(p)}. \tag{149}$$

Now let us consider the operational expansion suggested by the Heaviside processes:

$$\begin{aligned} x_s &= \frac{p}{\omega}\left(1+\frac{p^2}{\omega^2}\right)^{-1}\frac{1}{H(p)} \\ &= \left\{\frac{p}{\omega} - \left(\frac{p}{\omega}\right)^3 + \left(\frac{p}{\omega}\right)^5 - \left(\frac{p}{\omega}\right)^7 + \dots\right\}\frac{1}{(H(p)} \end{aligned} \tag{150}$$

and

$$\begin{aligned} x_c &= \left(\frac{p}{\omega}\right)^2\left(1+\frac{p^2}{\omega^2}\right)^{-1}\frac{1}{H(p)} \\ &= \left\{\left(\frac{p}{\omega}\right)^2 - \left(\frac{p}{\omega}\right)^4 + \left(\frac{p}{\omega}\right)^6 - \left(\frac{p}{\omega}\right)^8 + \dots\right\}\frac{1}{H(p)}. \end{aligned} \tag{151}$$

Now let us identify $1/H(p)$ with $h(t)$ and replace p^n by d^n/dt^n: we get

$$x_s = \left\{\frac{1}{\omega}\frac{d}{dt} - \frac{1}{\omega^3}\frac{d^3}{dt^3} + \frac{1}{\omega^5}\frac{d^5}{dt^5} - \dots\right\} h(t) \tag{152}$$

and

$$x_c = \left\{\frac{1}{\omega^2}\frac{d^2}{dt^2} - \frac{1}{\omega^4}\frac{d^4}{dt^4} + \frac{1}{\omega^6}\frac{d^6}{dt^6} - \dots\right\} h(t). \tag{153}$$

We have now to inquire into the significance of equations (152) and (153), derived from the operational equations of the response of the system of an e.m.f. $\sin \omega t$ and $\cos \omega t$ respectively, impressed at time $t=0$. From the mode of derivation of these expansions from the operational equations it might be inferred that they are the divergent of asymptotic expansions of the operational equations (147) and (149). This would certainly not be an unreasonable inference in the light of the Heaviside expansions we have just been considering. This inference is however, not correct: on the other hand, the series (152) and (153) have a definite physical significance, as we shall now show from the explicit equations of the problem.

By equation (31), the explicit equation for x_s, given operationally by (147), is

$$x_s = \frac{d}{dt}\int_0^t \sin \omega\tau . h(t-\tau)d\tau = \int_0^t \sin \omega(t-\tau)h'(\tau)d\tau + h(o) \sin \omega t \qquad (154)$$

where $h'(t) = d/dt\ h(t)$. By a well known trigonometric formula, this is

$$x_s = \sin \omega t \int_0^t \cos \omega t . h'(t)dt - \cos \omega t \int_0^t \sin \omega t . h'(t)dt + h(o) \sin \omega t.$$

Writing

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

this becomes

$$x_s = \sin \omega t \int_0^\infty \cos \omega t . h'(t)dt - \cos \omega t \int_0^\infty \sin \omega t . h'(t)dt$$

$$+ h(o) \sin \omega t - \int_t^\infty \sin \omega(t-\tau)h'(\tau)d\tau. \qquad (155)$$

The first three terms are simply the steady-state response to the impressed e.m.f. $\sin \omega t$: that is, they represent the ultimate steady state value of x_s when the transient oscillations have died away. The last term, which we shall denote by T_s, represents the transient oscillations which are set up when the e.m.f. is applied. Thus

$$T_s = -\int_t^\infty \sin \omega(t-\tau)h'(\tau)d\tau. \qquad (156)$$

Now from (156)

$$T_s = -\frac{1}{\omega}\int_t^\infty h'(\tau).d.\cos \omega(\tau - t)$$

and integrating by parts

$$T_s = \frac{1}{\omega}\frac{d}{dt}h(t) + \frac{1}{\omega}\int_t^\infty \cos \omega(\tau - t)\frac{d^2}{dt^2}h(\tau)d\tau. \qquad (157)$$

Repeating the process of partial integration, we get:

$$T_s = \frac{1}{\omega}\frac{d}{dt}h(t) - \frac{1}{\omega^2}\int_t^\infty \sin \omega(\tau - t)\frac{d^3}{d\tau^3}h(\tau)d\tau. \qquad (158)$$

Repeating the process again

$$T_s = \frac{1}{\omega}\frac{d}{dt}h(t) - \frac{1}{\omega^3}\frac{d^3}{dt^3}h(t) + \frac{1}{\omega^4}\int_t^\infty \sin \omega(\tau - t)\frac{d^5}{d\tau^5}h(\tau)d\tau.$$

This process can be repeated indefinitely, and we get

$$T_s=\left(\frac{1}{\omega}\frac{d}{dt}-\frac{1}{\omega^3}\frac{d^3}{dt^3}+\frac{1}{\omega^5}\frac{d^5}{dt^5}-\ldots+\frac{(-1)^{n-1}}{\omega^{2n-1}}\frac{d^{2n-1}}{dt^{2n-1}}\right)h(t)$$
$$+\frac{(-1)^n}{\omega^{2n}}\int_t^\infty \sin\omega(\tau-t)\frac{d^{2n+1}}{dt^{2n+1}}h(\tau)d\tau. \qquad (159)$$

The series expansion (159), except for the remainder term, is identical with the series expansion (152) derived directly from the operational equation. This series may be either convergent or divergent, depending on the frequency $\omega/2\pi$ and the character of the indicial admittance function $h(t)$. In the important problems of the building-up of alternating currents in cables and lines we shall see that, even when divergent, the series is of an asymptotic character and can be employed for computation.

We thus arrive at the following theorem:

If an e.m.f. $\sin\omega t$ is impressed at time $t=0$ on a network or system of generalized indicial admittance $h(t)$, and if the *transient distortion*, T_s, is defined as the instantaneous difference between the actual response of the system and the steady-state response, then T_s can be expressed as the series

$$\left(\frac{1}{\omega}\frac{d}{dt}-\frac{1}{\omega^3}\frac{d^3}{dt^3}+\frac{1}{\omega^5}\frac{d^5}{dt^5}-\ldots+\frac{(-1)^{n-1}}{\omega^{2n-1}}\frac{d^{2n-1}}{dt^{2n-1}}\right)h(t) \qquad (160)$$

with a remainder term

$$\frac{(-1)^n}{\omega^2}\int_t^\infty \sin\omega(\tau-t)\frac{d^{2n+1}}{dt^{2n+1}}h(\tau)d\tau.$$

If the impressed e.m.f. is $\cos\omega t$, the corresponding series for the transient distortion, T_c, is

$$\left(\frac{1}{\omega^2}\frac{d^2}{dt^2}-\frac{1}{\omega_4}\frac{d_4}{dt_4}+\frac{1}{\omega^6}\frac{d^6}{dt^6}-\ldots-\frac{(-1)^n}{\omega^{2n}}\frac{d^{2n}}{dt^{2n}}\right)h(t) \qquad (161)$$

with a remainder term

$$\frac{(-1)^n}{\omega^{2n}}\int_t^\infty \cos\omega(\tau-t)\frac{d^{2n+1}}{dt^{2n+1}}h(\tau)d\tau.$$

The second part of this theorem, relating to the transient distortion, T_c, in response to an e.m.f. $\cos\omega t$, is derived from formula (31) by processes precisely analagous to those employed above in deriving the series expansion for T_s. The derivation will be left to the reader.

To summarize the preceding discussion of the divergent solution of operational equations, it may be said that the theory is as yet rather unsatisfactory. To the physicist it is unsatisfactory because he requires an automatic rule giving a correct asymptotic expansion by purely algebraic operations without investigations of remainder terms or auxiliary functions. Furthermore, the precise sense in which the expansion asymptotically represents the solution cannot be stated in general, but requires an independent investigation in the case of each individual problem.

On the other hand when an asymptotic expansion is known to exist, the Heaviside Rule finds this expansion with incomparable directness and simplicity, the problem of justifying the expansion being a purely mathematical one, which usually need not trouble the physicist. Furthermore, on the purely mathematical side, the Heaviside Rule is of large interest and should lead to interesting developments in the theory of asymptotic expansions.

(*To be continued*)

Abstracts of Bell System Technical Papers Not Appearing in this Journal

Commercial Loading of Telephone Cables. W. FONDILLER.[1] The application of loading coils to exchange area cable and to toll cable is discussed and data given on the loading coils and the transmission characteristics of loaded cable circuits.

An important section of the paper deals with the requirements for loading phantom circuits. In particular, the crosstalk and noise requirements for phantom loading are analyzed.

The paper concludes with a comparative study of three systems of phantom loading which are in commercial use, viz., the Campbell-Shaw, the Ebling and the Olsen-Pleijel system. It is concluded that the Campbell-Shaw phantom loading system, which has been adopted as standard by the Bell System, as well as by many European Administrations (notably the British Post Office), has marked advantages over the other two systems which have been used to a minor extent in continental Europe.

The Schottky Effect in Low Frequency Circuits,[2] by J. B. Johnson. This effect, discovered by Schottky, which depends on the probability of fluctuations of electron emission from a filament, has been measured over a considerable range of conditions in resonant circuits of which the natural frequency was varied from 8 to nearly 6000 p.p.s. The effect is much larger in the lower range of frequencies than the theory predicts. With a tungsten filament, the ratio of observed to theoretical effect e'/e is about .7 for frequencies above 200, but increases rapidly to 50 at 10 cycles per sec. With an oxide coated filament, the ratio increases from 1 at 5000 cycles to 100 at 100 cycles. This is interpreted to mean that the emission of electrons is not strictly chaotic but is influenced by irregular temporal changes in the cathode emissivity. In a high frequency circuit these changes become imperceptible and the emission is effectively random. When current is limited by space charge the Schottky effect decreases because of the interaction of the electrons, and other disturbances may act upon the space charge so as to completely mask the remanent Schottky effect. The magnitude of the disturbances in amplifying vacuum tubes can therefore not be predicted from measurements on the true Schottky effect.

A Note on Schottky's Method of Determining the Distribution of Velocities Among Thermionic Electrons,[3] C. Davisson. Limiting con-

[1] Electrical Communication, July, 1925.

[2] Physical Review, Vol. 26, No. 1, page 71, July, 1925.

[3] Physical Review, Vol. 25, No. 6, page 808, June, 1925.

ditions for Schottky's formula for the thermionic current from a filament to a coaxial cylinder. The formula must fail when, due to space charge, the potential at any distance $x(r < x < R)$ from the axis is less than $Vr^2 (R^2-x^2)/x^2(R^2-r^2)$, V being the potential of the filament with respect to the cylinder, and r and R the radii of filament and cylinder respectively. This is more restrictive than the condition for failure which has been previously assumed.

Variation of the Photo-electric Effect with Temperature in the Alkali Metals,[4] Herbert E. Ives and A. L. Johnsrud. Special cells having a hollow central cathode were immersed in liquid air for an extended period to condense any gases present on the outer alkali metal coated walls. By a stream of evaporating liquid air, the temperature of the cathode was held at temperatures between +20 and —180°C. In these cells the variation of photo-electric current with temperature in sodium, potassium and rubidium is continuous. The effect is relatively small for sodium, showing hardly at all for blue light or white light, but clearly for yellow light. The behavior of rubidium is similar to that previously reported for potassium. In a second form of cell, potassium was collected in a deep pool. By slowly cooling the metal from the molten conditions, smooth crystalline surfaces were obtained. With these annealed potassium surfaces, the variation of photo-electric current with temperature is represented by curves varying systematically in shape with the color of the light, and the effect is far greater than previously reported, amounting, for yellow light, to a variation of 10 to 15 times between room and liquid air temperature. When the surface is roughened curves of the previously reported type are obtained. Small pools give erratic effects, showing changes in opposite directions for different portions of the temperature range. It is concluded that the variation of photo-electric effect is intimately connected with the strains produced in the surface by expansion and contraction with temperature.

Echo Suppressors for Long Telephone Circuits,[5] A. B. Clark and R. C. Mathes. A device has been developed by the Bell System for suppressing "echo" effects which may be encountered under certain conditions in telephone circuits which are electrically very long. This device has been given the name "echo suppressor" and consists of relays in combination with vacuum tubes, which are operated by the voice currents so as to block the echoes without disturbing the main transmission.

[4] Physical Review, Vol. 25, No. 6, page 893, June, 1925.

[5] Jour. A. I. E. E., Vol. XLIV, No. 6, page 618, June, 1925.

This paper gives a brief description of this device, together with a discussion of its possibilities and limitations. A number of echo suppressors have been operated on commercial telephone circuits for a considerable period so that their practicability has been demonstrated.

Recent Commercial Development in Short Wave Transmitters and Receivers.[6] S. E. ANDERSON, L. M. CLEMENT, and G. C. DECOUTOULY. This paper describes the transmitter and receiver recently developed for use by the United States Coast Guard. This apparatus is for operation on wave lengths between 100 and 200 meters. In describing the development of the transmitter a short summary of the various circuit considerations is included. The actual transmitter finally developed is also described together with its operating characteristics.

In considering the radio receiver the various problems to be met in the design of a radio receiver of this character are dealt with at some length. The frequency characteristics of the radio receiver, as developed, are shown, and the method of determining them is described in detail.

The transmitter and receiver performed very satisfactorily under conditions more severe than will be met in actual service.

The Distribution of Initial Velocities Among Thermionic Electrons.[7] L. H. GERMER. The method used was to measure the number of electrons from a straight tungsten filament which were able to arrive at a co-axial cylindrical electrode against various retarding potentials. In order to eliminate certain disturbing factors, particularly photoelectric effects, this electrode was made in the form of a very fine grid and those electrons passing between the grid wires were collected upon an outside electrode and there measured. A rather complicated intermittent heating current arrangement allowed emission from the filament only when its surface was at uniform potential, and insured that the retarding potential had exactly the desired value. A current regulator kept the heating current constant to 1/30 per cent.

Electrons from Tungsten. Measurements of the variation of electron current with voltage were made at eight different temperatures ranging from 1440°K to 2475°K. Correction was made for the contact potential difference between filament and grid. At each temperature it was found that, except in the range of voltage where the current was limited by the space charge phenomenon, the current varied with voltage in just the manner calculated upon the assumption that the electrons leave the filament with velocity components distributed according to Maxwell's law for an electron atmosphere in temperature

[6] Proc. of I. R. E., Vol. 13, No. 4, page 413, August, 1925.

[7] Physical Review, Vol. 25, No. 6, page 795, June, 1925.

equilibrium with the hot filament. At 2175°K the assumed Maxwell distribution was verified up to a retarding potential so great that only one electron out of 10^{10} emitted electrons was able to reach the collector. It is believed that the present results are more reliable and extensive than any hitherto obtained, and that they are conclusive for electron emission from tungtsen in a high vacuum.

Electrons from Oxide Coated Platinum. Subsequent measurements by Dr. C. Davisson have shown that the electrons emitted from Wehnelt cathodes also have velocity components distributed according to Maxwell's law.

Automobile-Noise Measurement.[3] H. Clyde Snook. Automobile noise, although useful as a detector of mechanical imperfections of car operation, is otherwise so extremely undesirable that elaborate methods for analysis with a view toward preventing or suppressing such noise are warranted. The author presents an illustrated and detailed description of the mechanism of human hearing, according to studies made in the interests of telephonic transmission of maximum effectiveness, enumerating and explaining the devices developed for evaluating the sources of sound and its modes of propagation and amplification.

An automobile can be considered to be composed of a number of acoustic resonators having varied degrees of coupling between them, and comparisons are made of the velocity of sound propagation through the different materials with that of its transmission in air, the velocity being greater in the structural material. The apparatus used for the detection of noise and its measurement consists of varied types of equipment, divided into two classes; one includes the contact type and the other the air-impact type, both being demonstrated.

Following an enumeration of the different detectors and auxiliary apparatus in use and comments upon the methods employed, it is stated among other conclusions that it seems advisable to base loudness measurements of automobile noise upon the difference of energy between the measured sound and an arbitrary standard of sound which is the threshold of normal hearing; that, to locate the origin of automobile noise, it frequently is sufficient merely to detect the noise without measuring its loudness; and that, to identify the origin of automobile noise, it often is of value to ascertain its component frequencies.

[3] Jour. Soc. of Automotive Engineers, Vol. XVII, No. 1, page 115, July, 1925.

Contributors to this Issue

H. P. Charlesworth, B.S., Massachusetts Institute of Technology, 1905; Engineering Department, American Telephone and Telegraph Company, 1905–19; Equipment and Transmission Engineer, Department of Operation and Engineering, 1919; Plant Engineer, 1920—. Mr. Charlesworth has had broad experience in the development of telephone equipment and with traffic conditions and the standardization of operating methods and practices.

G. A. Pennock, B.S., Massachusetts Institute of Technology, 1899; Secretary, Kansas City Bolt & Nut Company, 1899–1901; Chief Draftsman, Weber Gas & Gasoline Engineering Company, Kansas City, Missouri, 1901–1902; Mechanical Superintendent, Rock Island Plow Company, 1902–1906; with Western Electric Company from 1906, as Factory Engineer, European Plant Engineer and Technical Superintendent.

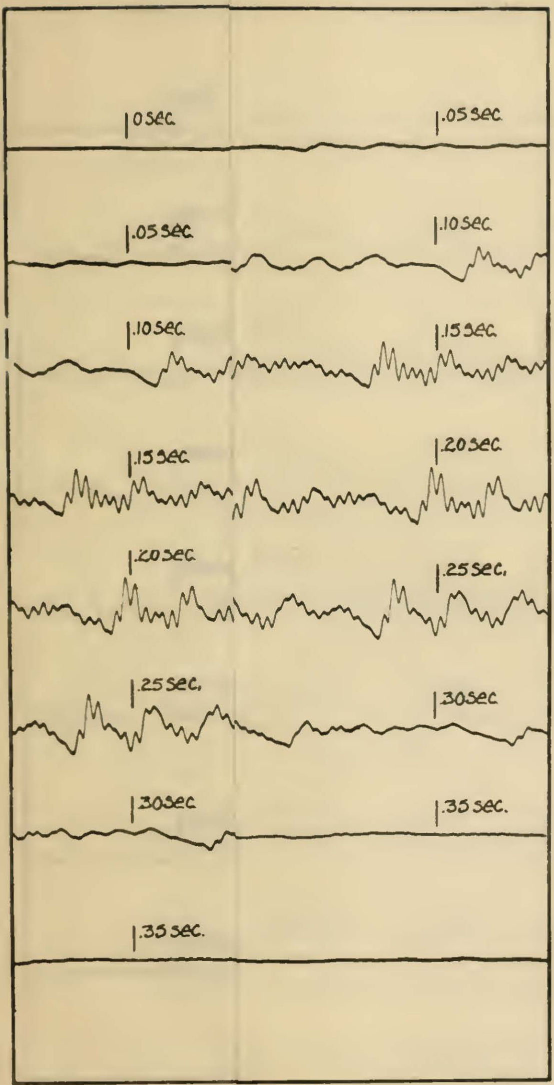
George Crisson, M.E., Stevens Institute of Technology, 1906; instructor in Electrical Engineering, 1906–10. American Telephone and Telegraph Company, Engineering Department, outside plant division, 1910–14; transmission and protection division, 1914–19; Development and Research Department, transmission development division, 1919 —.

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, 1913–24; Bell Telephone Laboratories, Inc., 1925 —. 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.

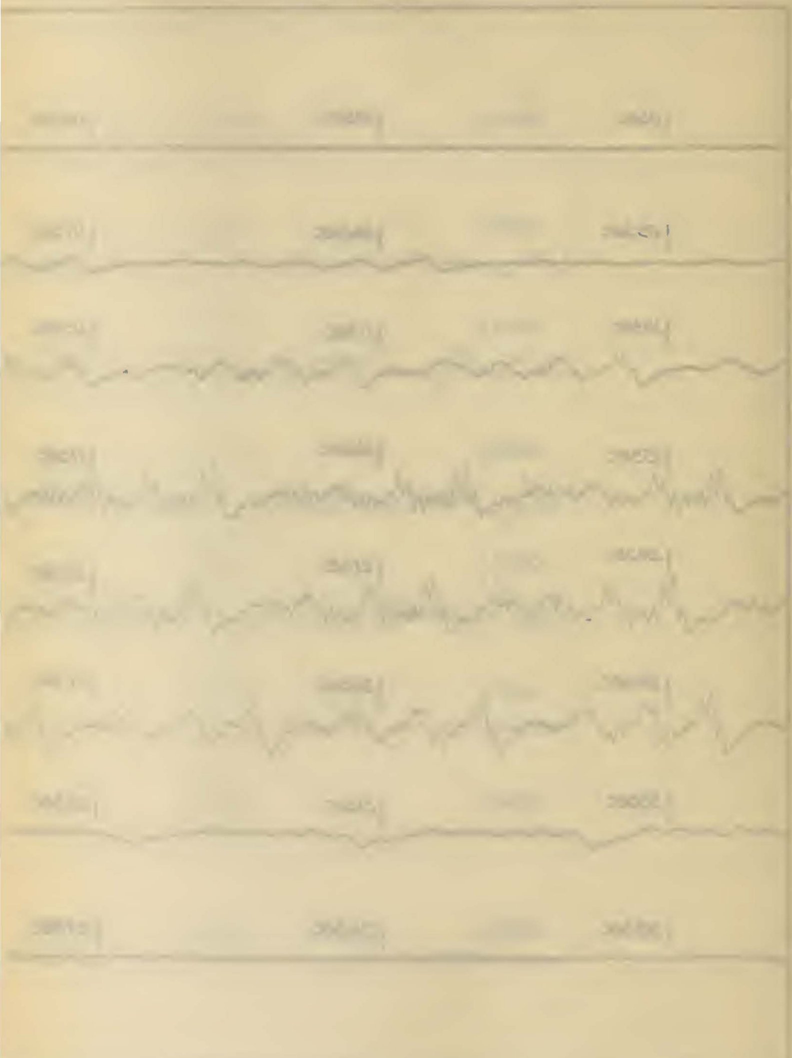
C. F. Sacia, B.E.E., University of Michigan, 1916; Engineering Department of the Western Electric Company, 1916–24; Bell Telephone Laboratories, Inc., 1925—. Mr. Sacia has been engaged upon methods for recording and analysing speech.

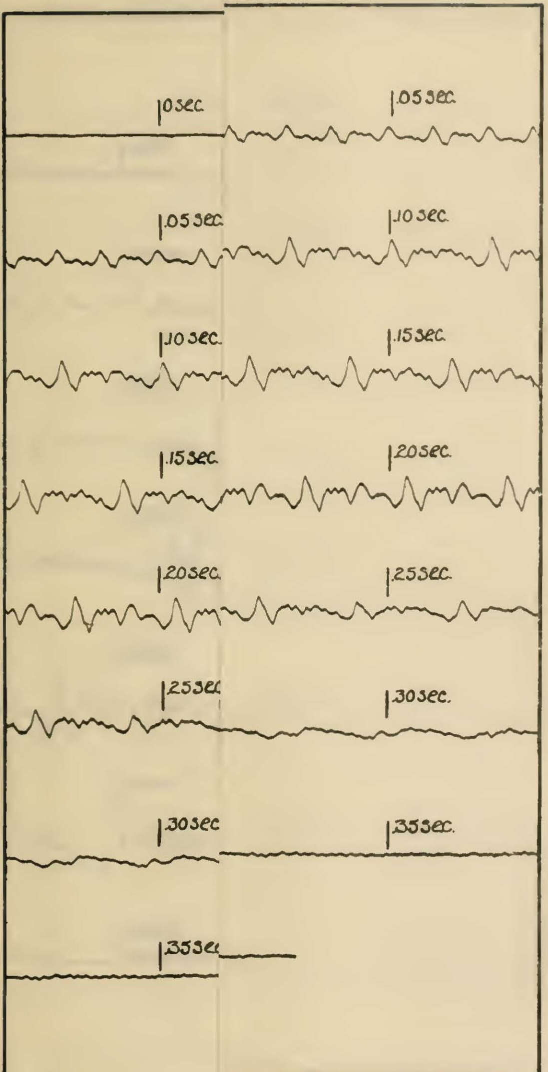
Karl K. Darrow, S.B., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., in physics and mathematics, University of Chicago, 1917; Engineering Department Western Electric Company, 1917–24; Bell Telephone Laboratories, Inc., 1925–. Mr. Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

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.

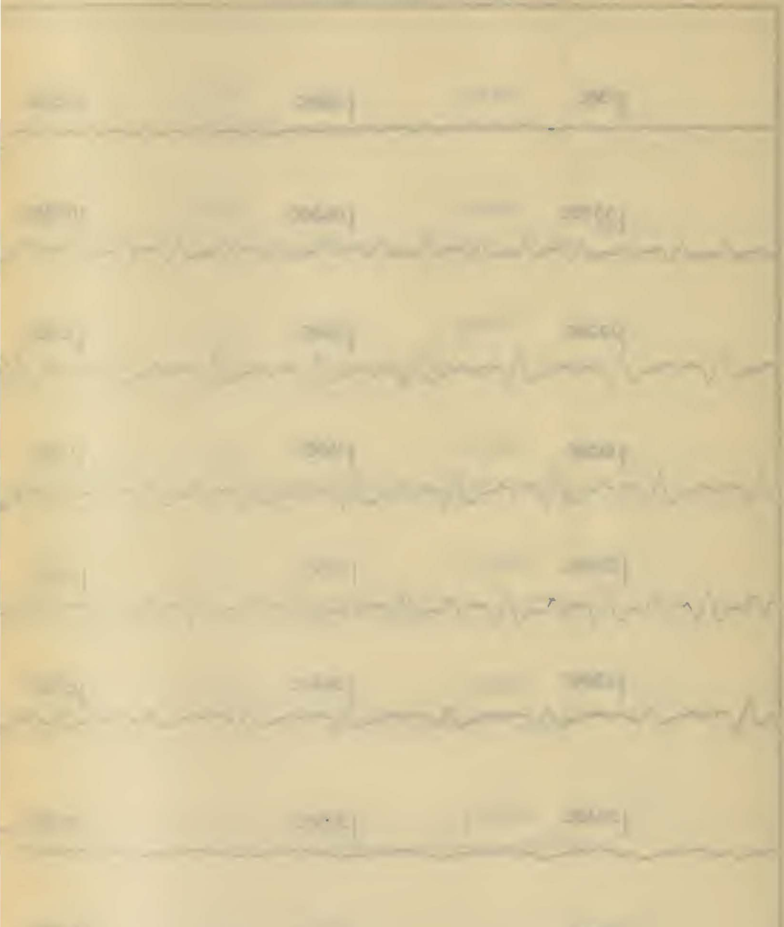

0 sec.
.05 sec.
.05 sec.
.10 sec.
.10 sec.
.15 sec.
.15 sec.
.20 sec.
.20 sec.
.25 sec.
.25 sec.
.30 sec.
.30 sec.
.35 sec.
.35 sec.

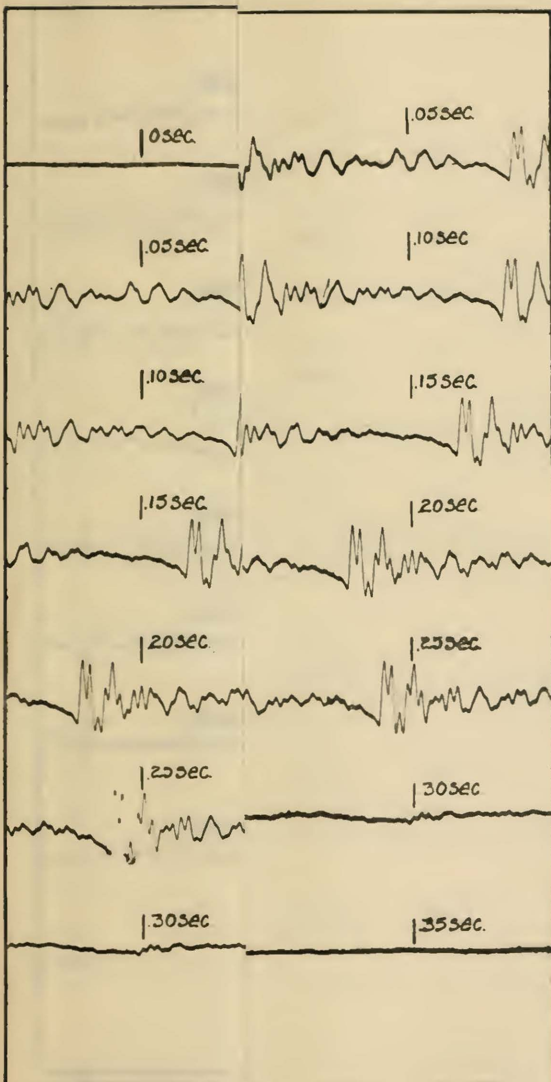
(1)

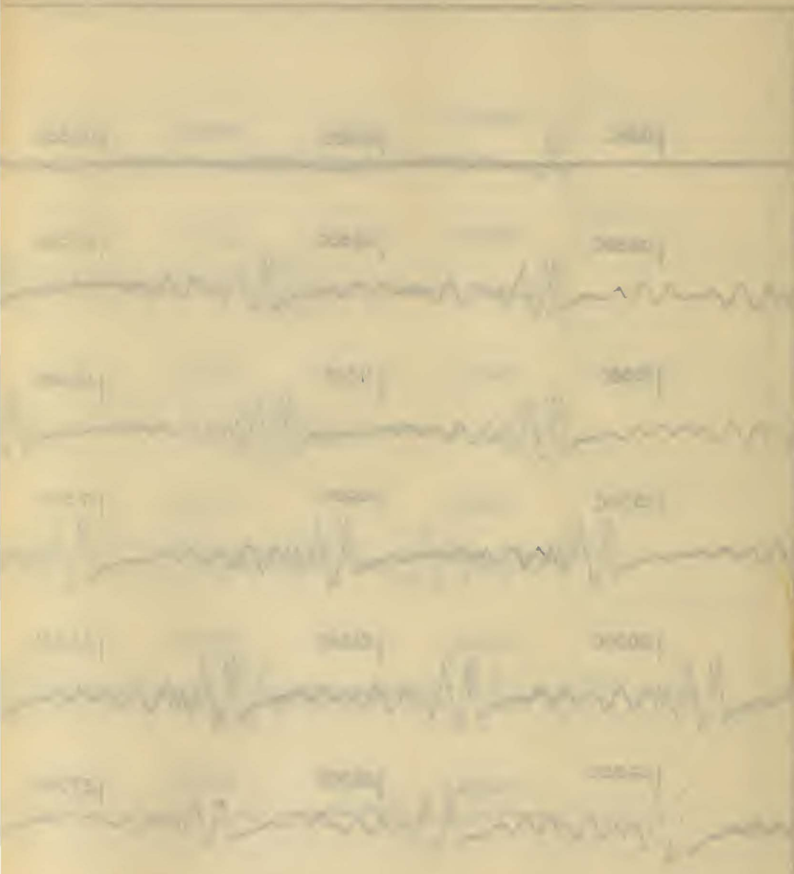


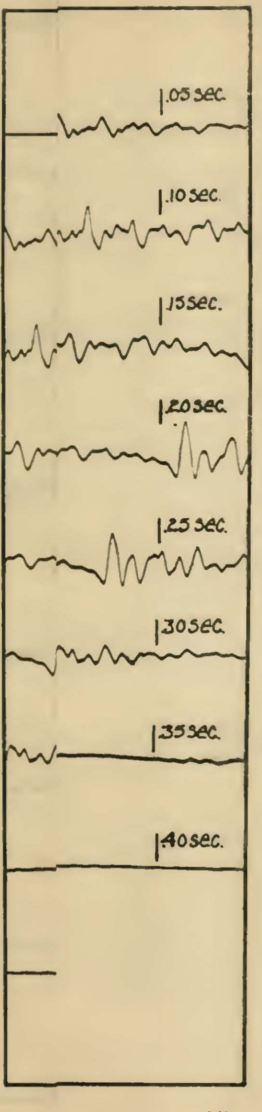

0 sec.
.05 sec.
.05 sec.
.10 sec.
.10 sec.
.15 sec.
.15 sec.
.20 sec.
.20 sec.
.25 sec.
.25 sec
.30 sec.
.30 sec
.35 sec.
.35 sec

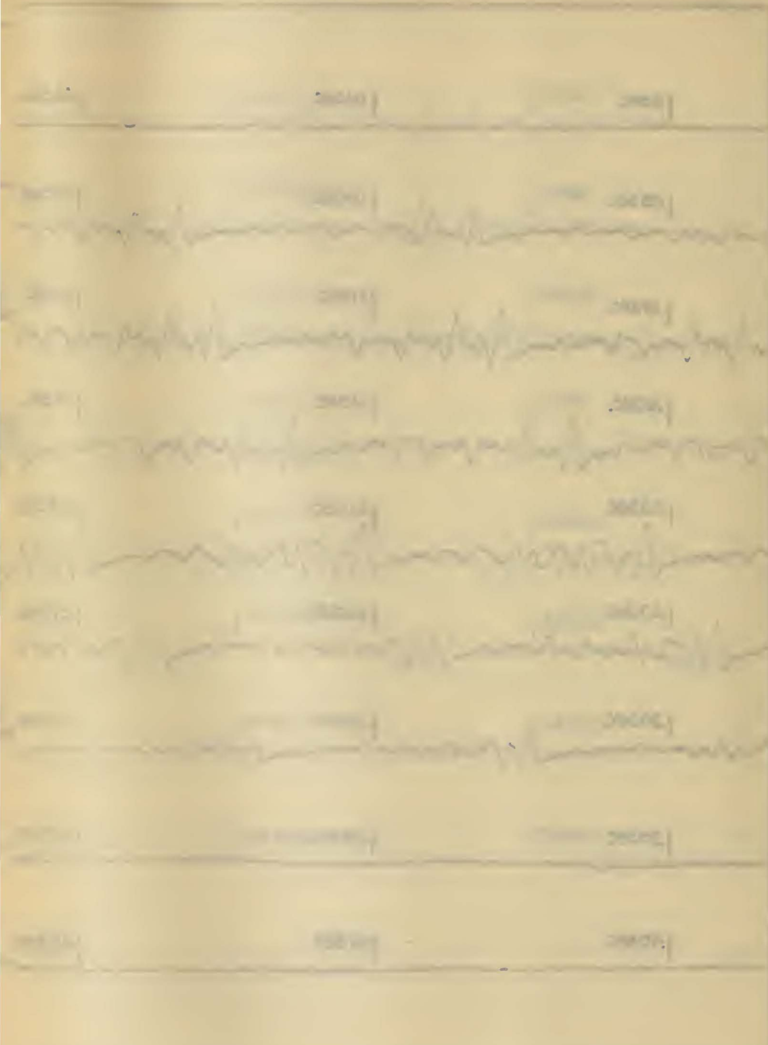
(2)

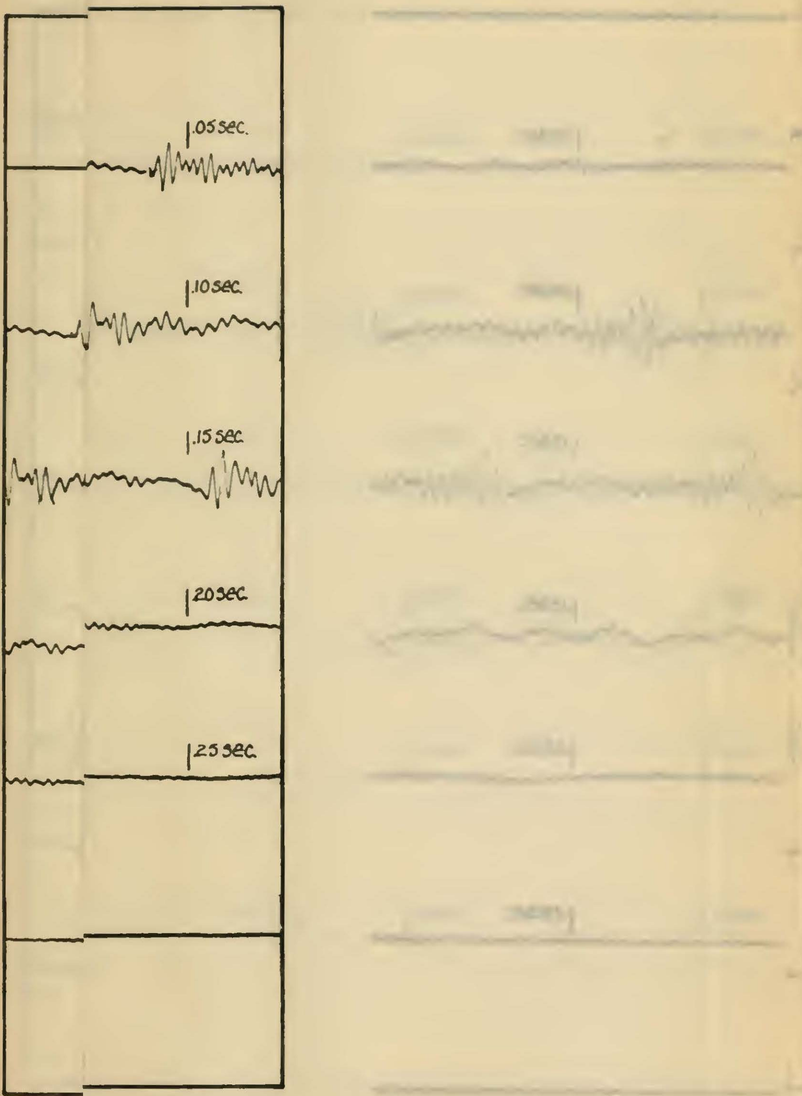


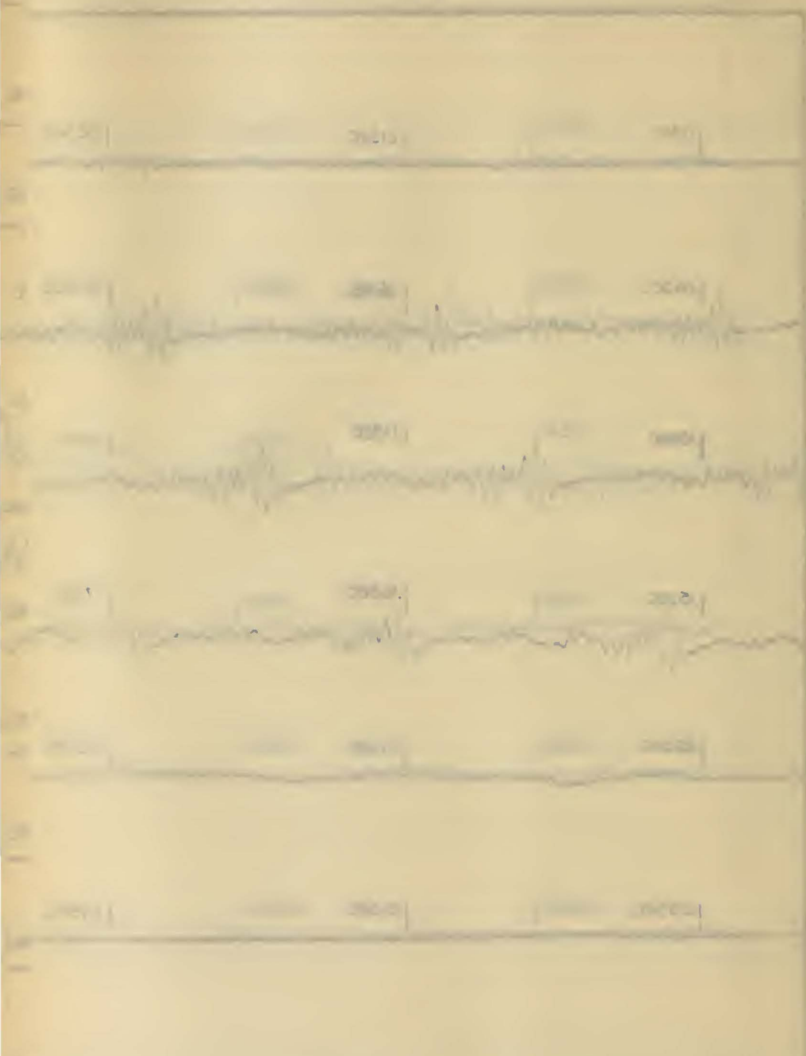

0 sec.
.05 sec.
.05 sec.
.10 sec
.10 sec.
.15 sec.
.15 sec.
.20 sec
.20 sec.
.25 sec.
.25 sec.
.30 sec
.30 sec
.35 sec.

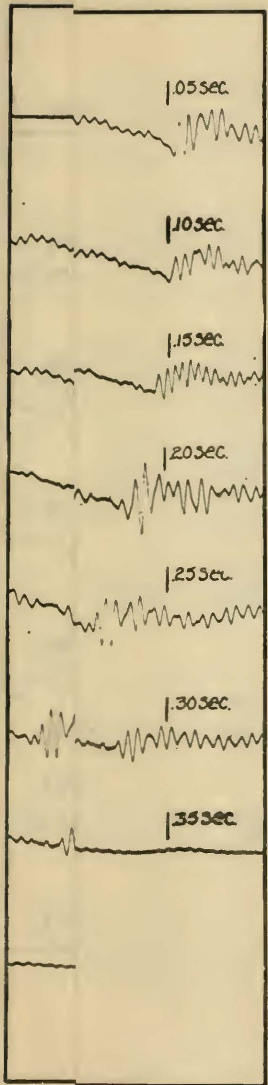


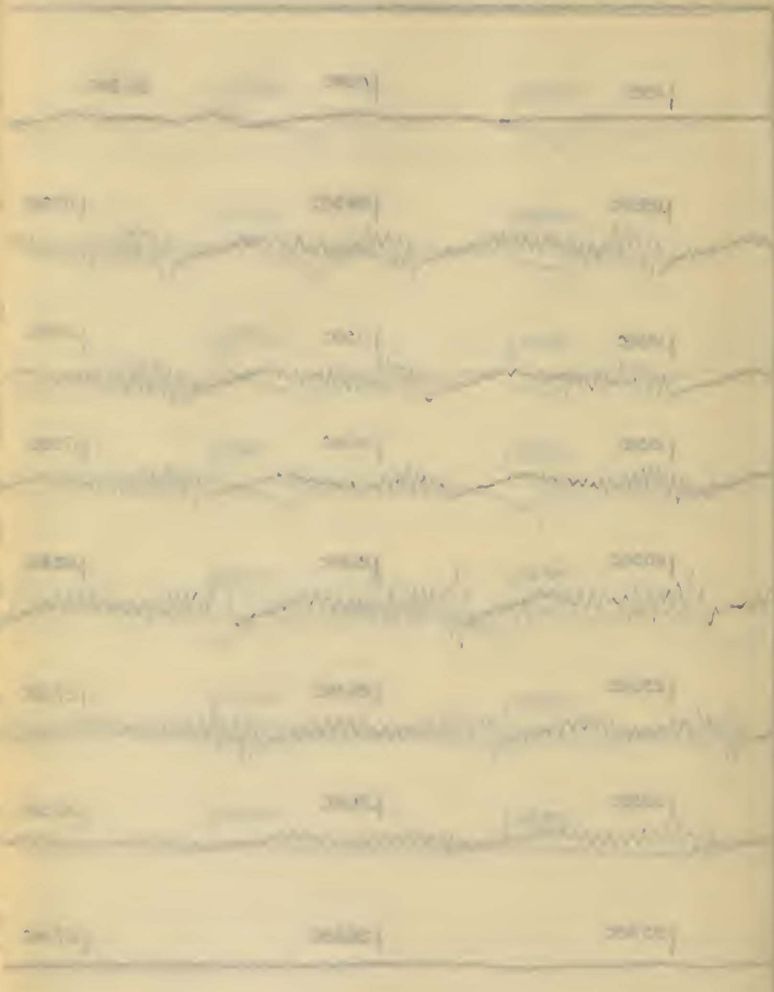

.05 sec.
.10 sec.
.15 sec.
.20 sec.
.25 sec.
.30 sec.
.35 sec.
.40 sec.

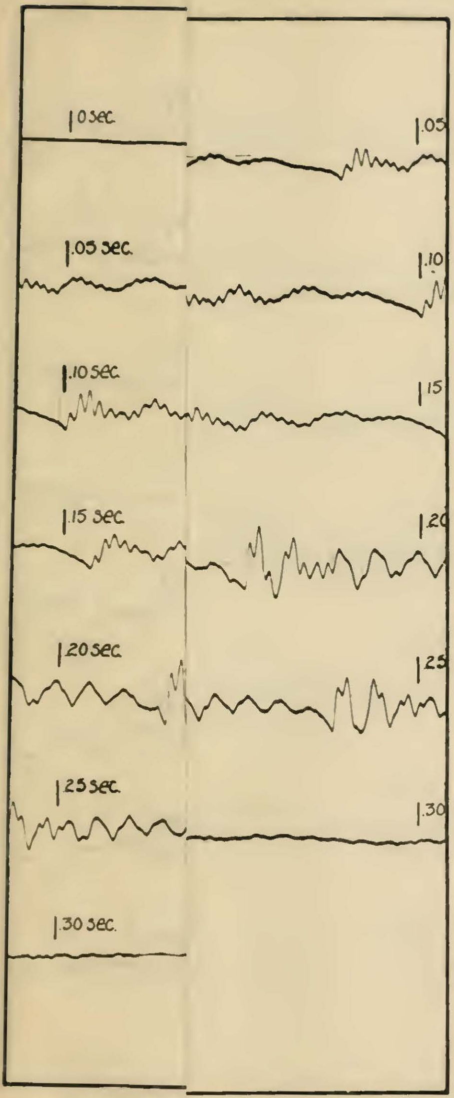


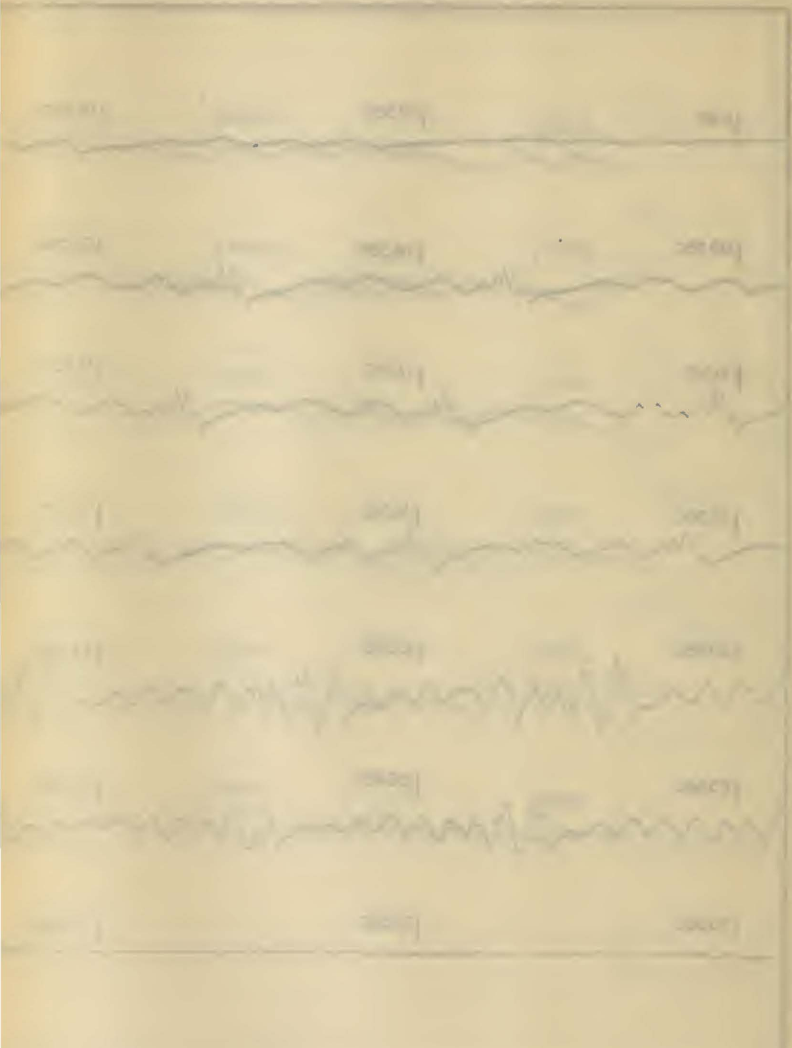

.05 sec.
.10 sec.
.15 sec.
.20 sec.
.25 sec.

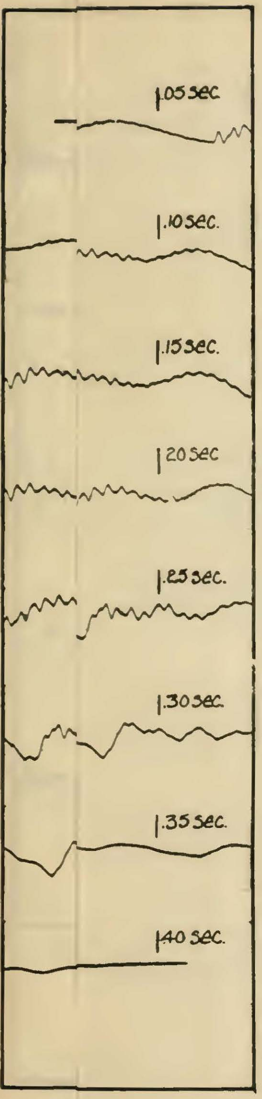


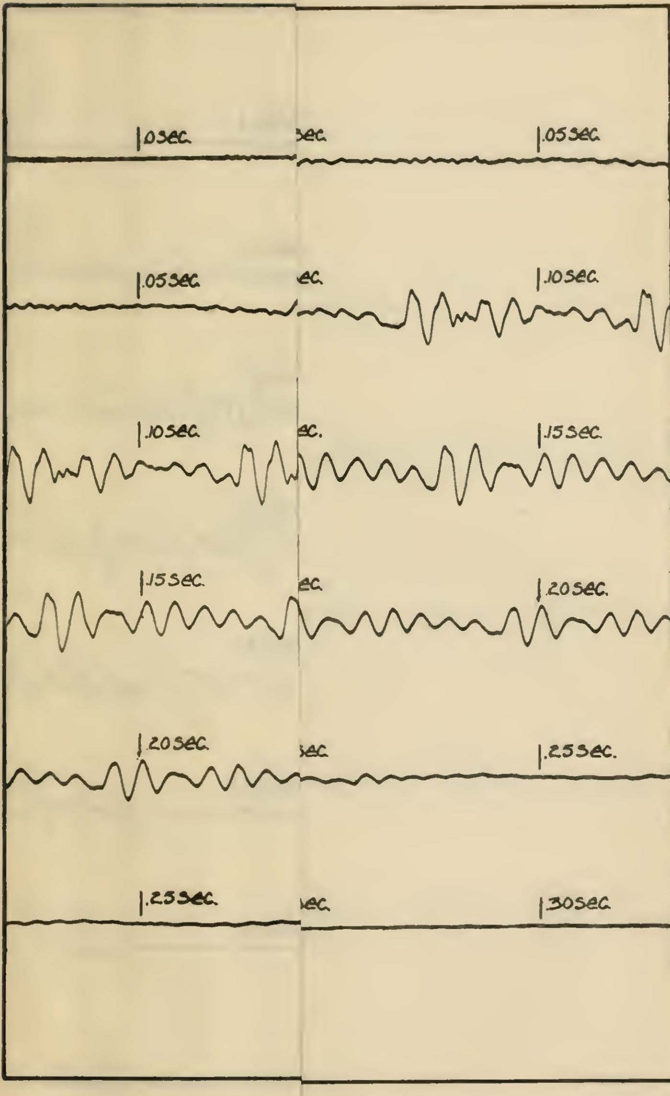

.05 sec.
.10 sec.
.15 sec.
.20 sec.
.25 sec.
.30 sec.
.35 sec.

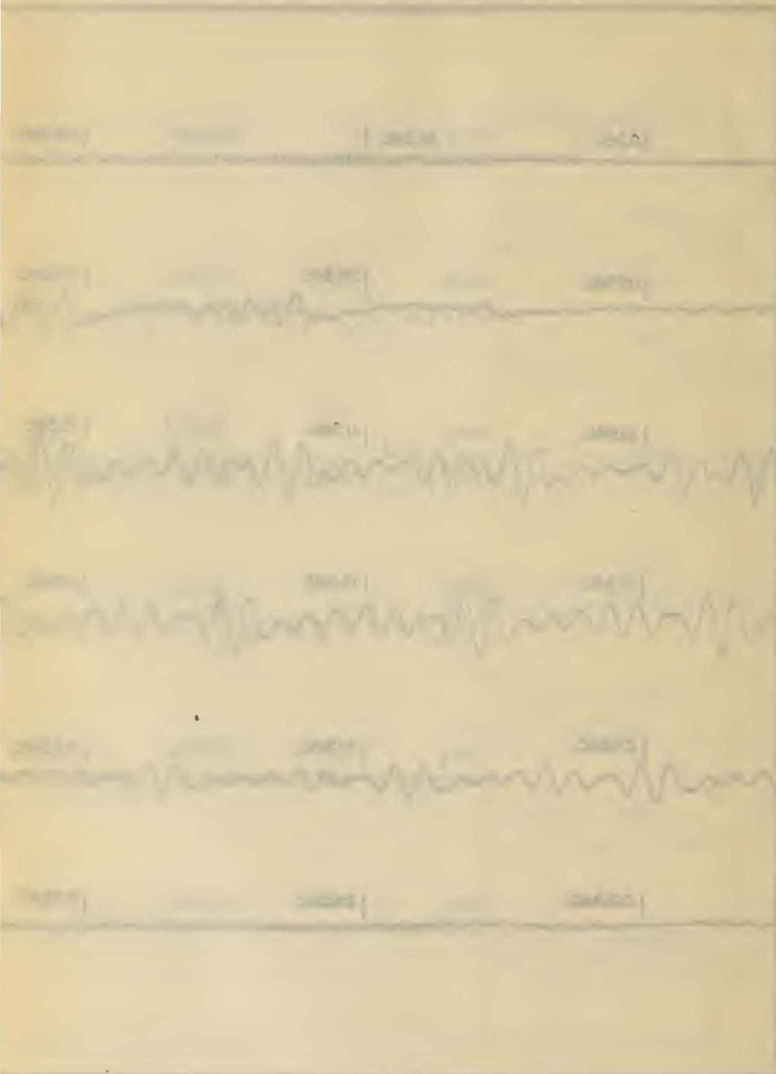


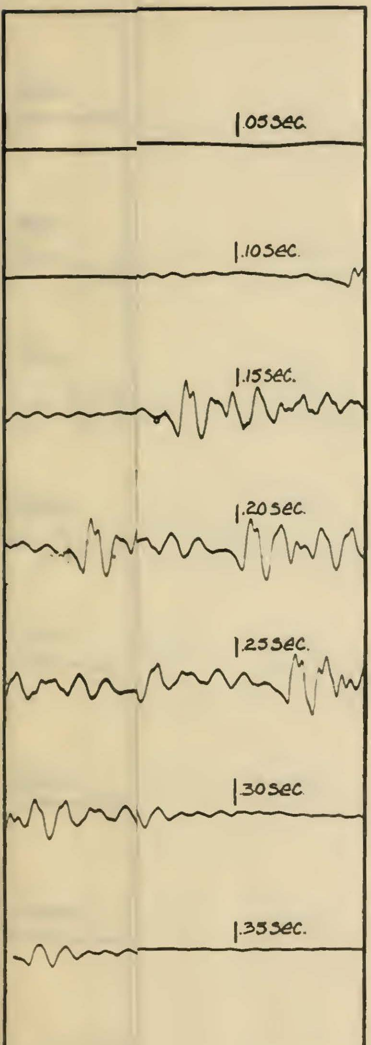

0 sec.
.05
.05 sec.
.10
.10 sec.
.15
.15 sec.
.20
.20 sec.
.25
.25 sec.
.30
.30 sec.

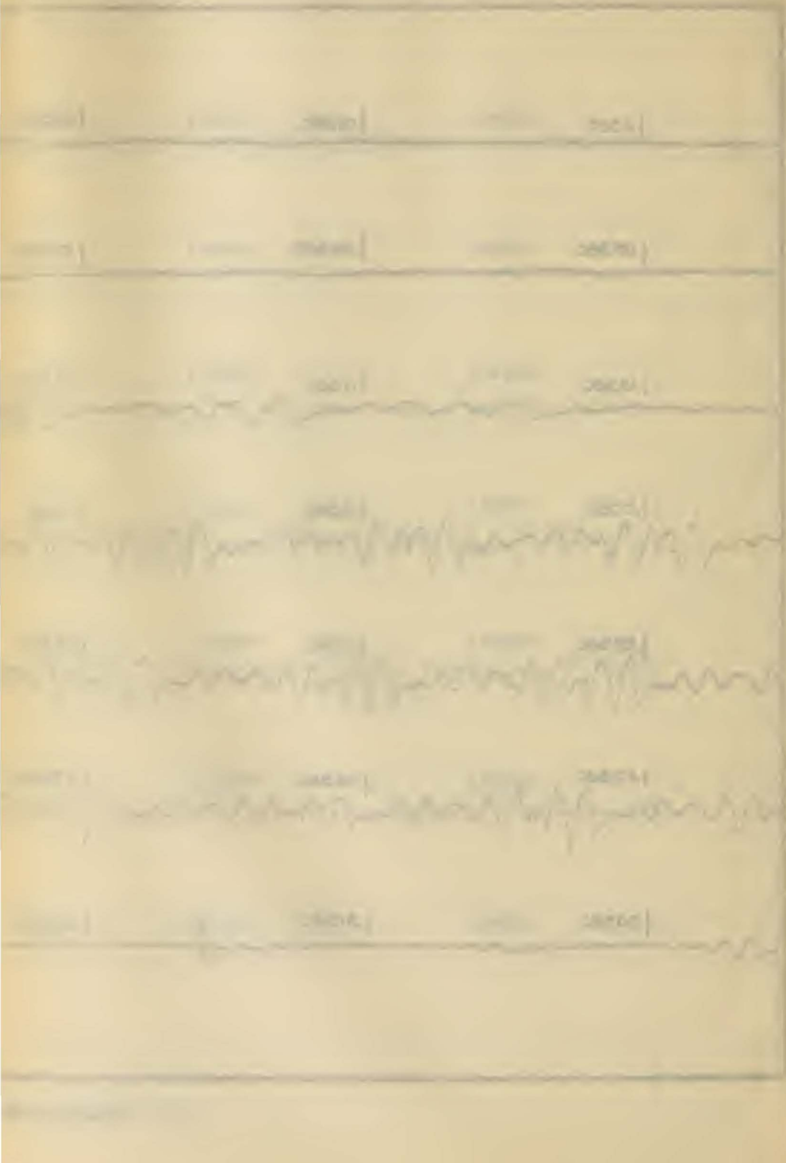


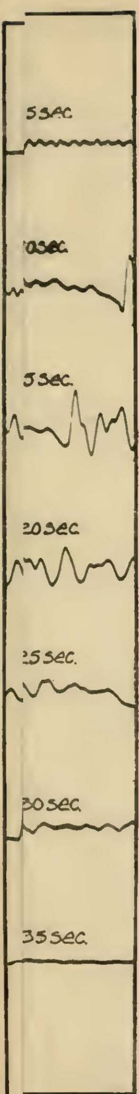

.05 sec.
.10 sec.
.15 sec.
.20 sec.
.25 sec.
.30 sec.
.35 sec.
.40 sec.


.0 sec.
sec.
.05 sec.
.05 sec.
ec.
.10 sec.
.10 sec.
ec.
.15 sec.
.15 sec.
ec.
.20 sec.
.20 sec.
sec.
.25 sec.
.25 sec.
sec.
.30 sec.

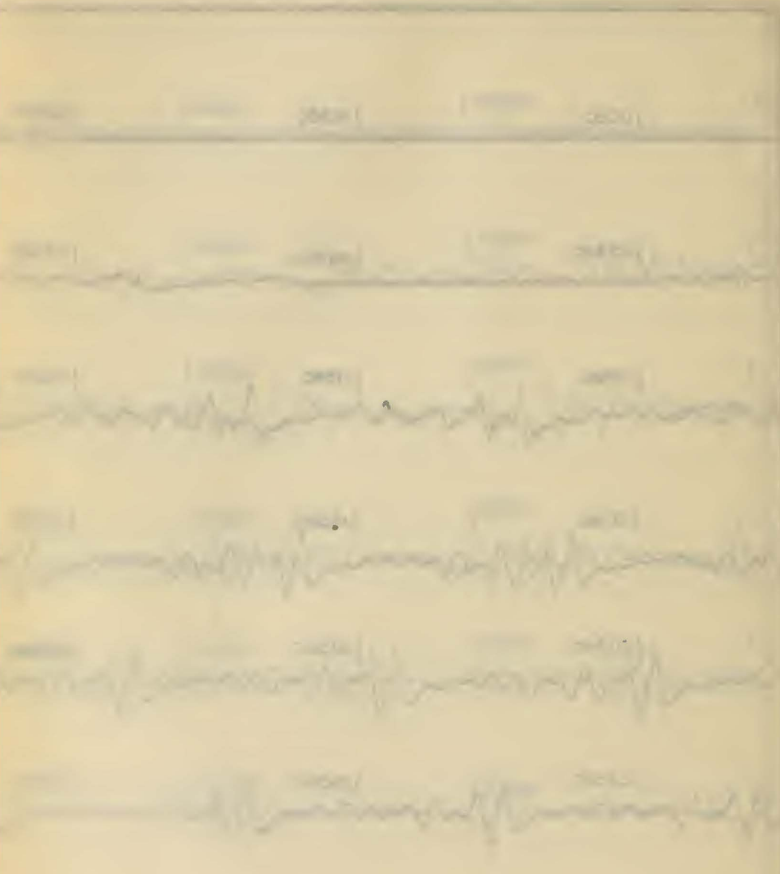


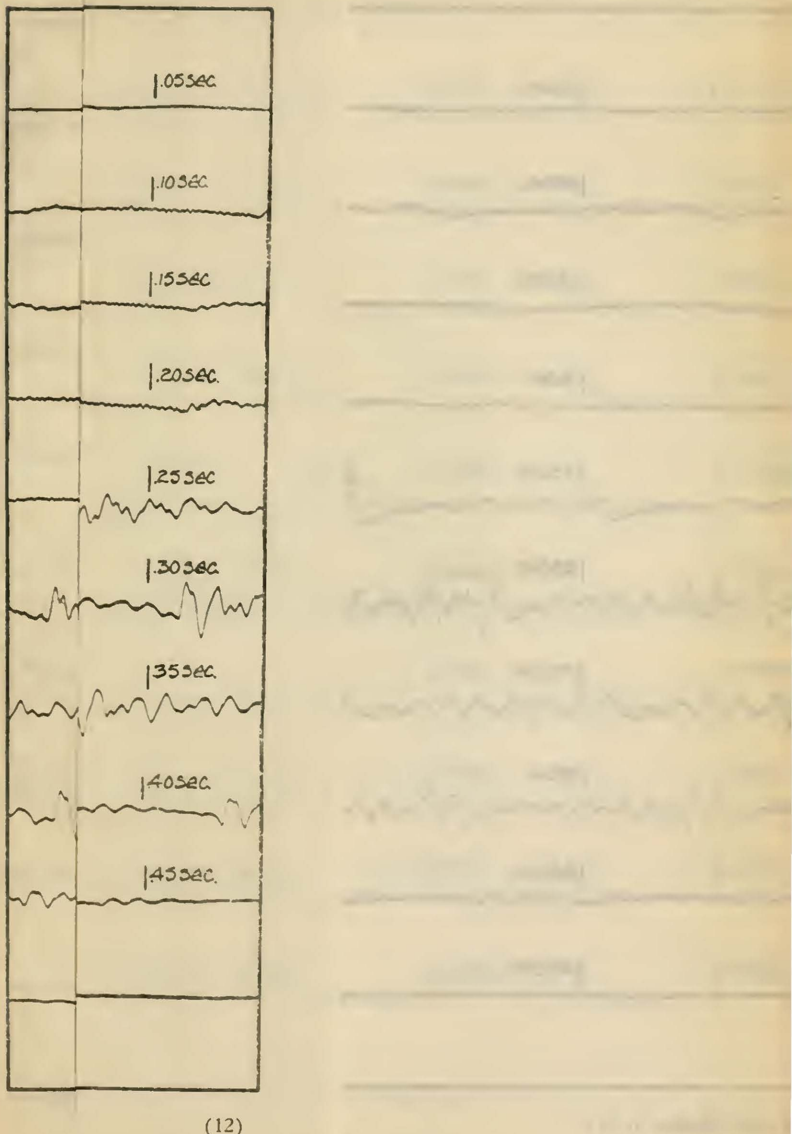

.05 sec.
.10 sec.
.15 sec.
.20 sec.
.25 sec.
.30 sec.
.35 sec.

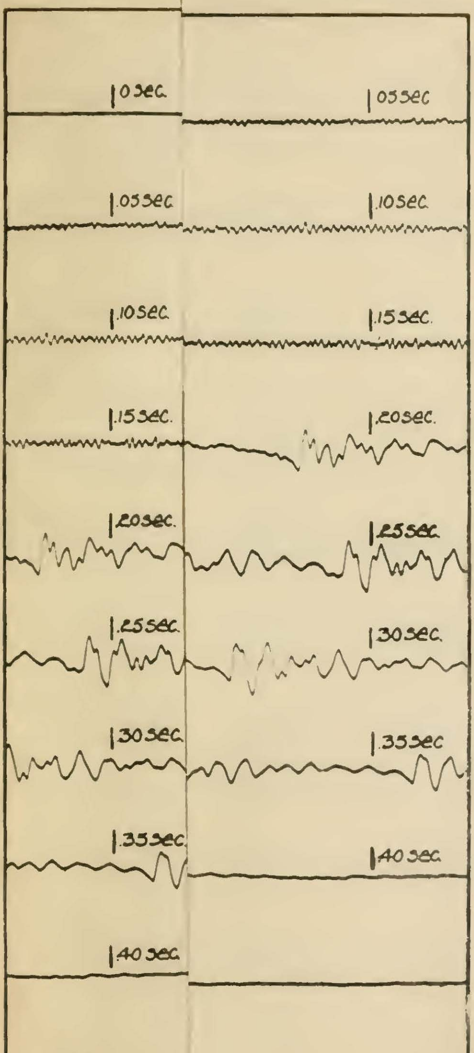




5 sec.
0 sec.
5 sec.
20 sec.
25 sec.
30 sec.
35 sec.




.05 sec.
.10 sec.
.15 sec.
.20 sec.
.25 sec.
.30 sec.
.35 sec.
.40 sec.
.45 sec.



0 sec.
.05 sec
.05 sec.
.10 sec.
.10 sec.
.15 sec.
.15 sec.
.20 sec.
.20 sec.
.25 sec.
.25 sec.
.30 sec.
.30 sec.
.35 sec
.35 sec.
.40 sec.
.40 sec.

Index to Volume IV

A

B

C

D

E

F

G

H

I

J

L

M

N

O

P

R

S

T

W