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.

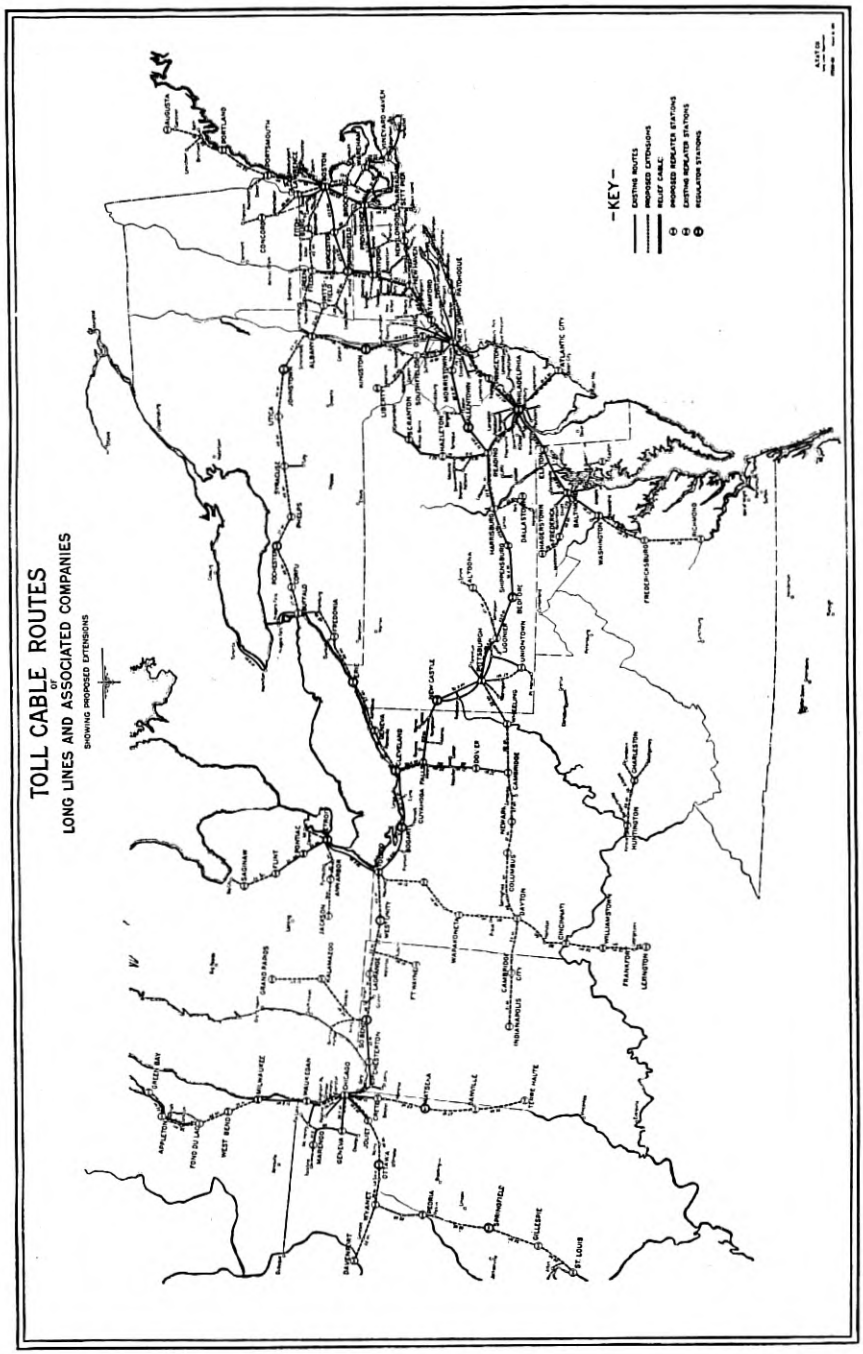

TOLL CABLE ROUTES
OF
LONG LINES AND ASSOCIATED COMPANIES
SHOWING PROPOSED EXTENSIONS
-KEY-
EXISTING ROUTES
PROPOSED EXTENSIONS
RELIEF CABLE
PROPOSED REPEATER STATIONS
EXISTING REPEATER STATIONS
TERMINAL STATIONS

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

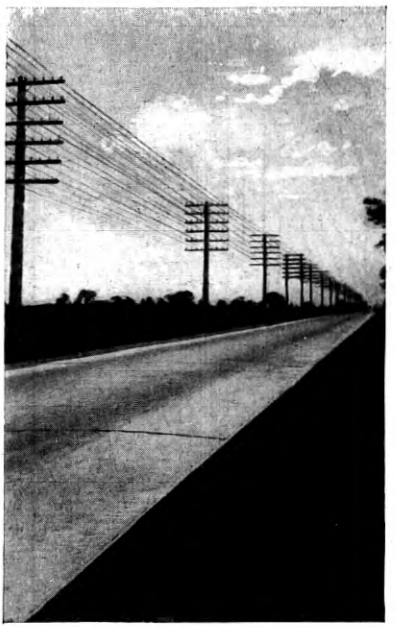


Fig. 2—Open wire toll line

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

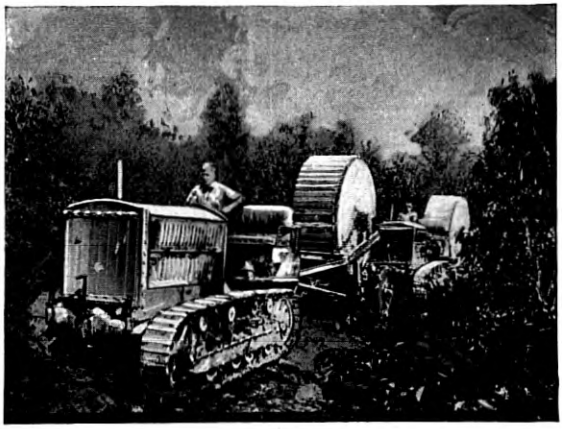


Fig. 3—Transporting cable reels through Allegheny Mountains

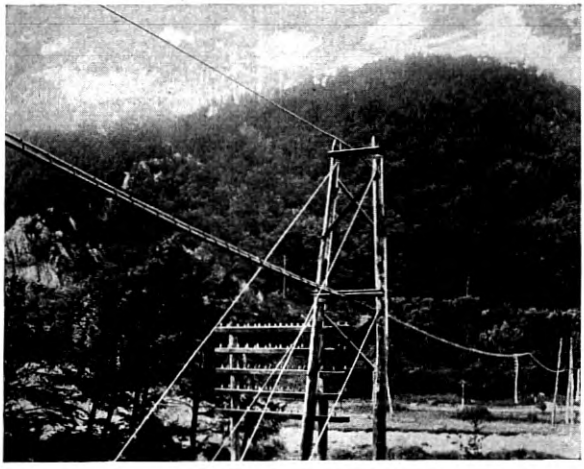


Fig. 4—Toll cable line in Allegheny Mountains

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

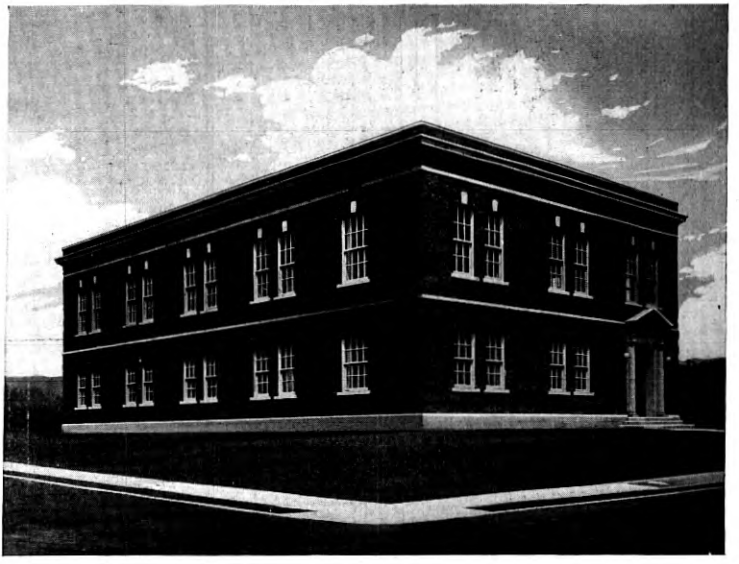


Fig. 5—Typical telephone repeater station

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

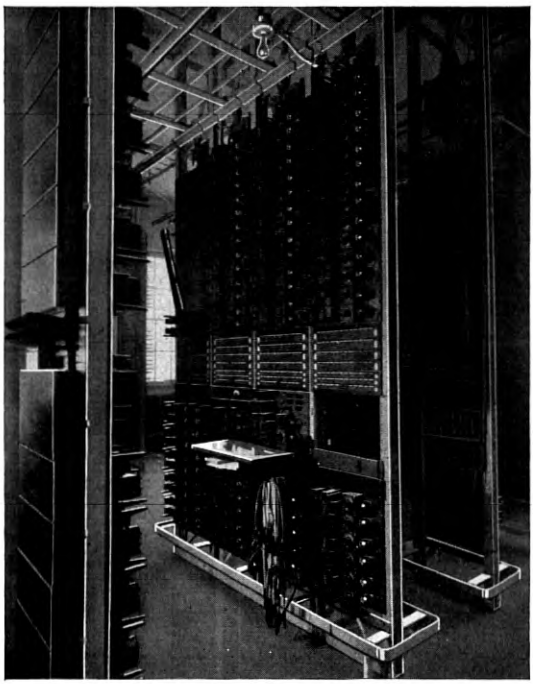


Fig. 6—Bank of 2-wire telephone repeaters

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

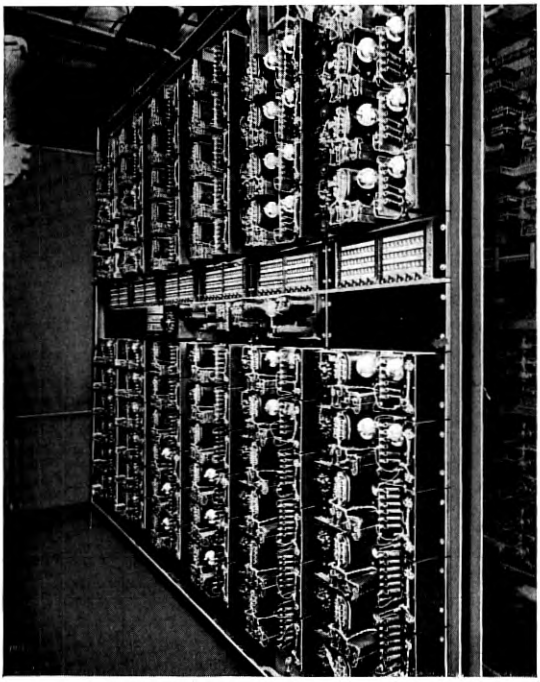


Fig. 7—Bank of 4-wire telephone repeaters

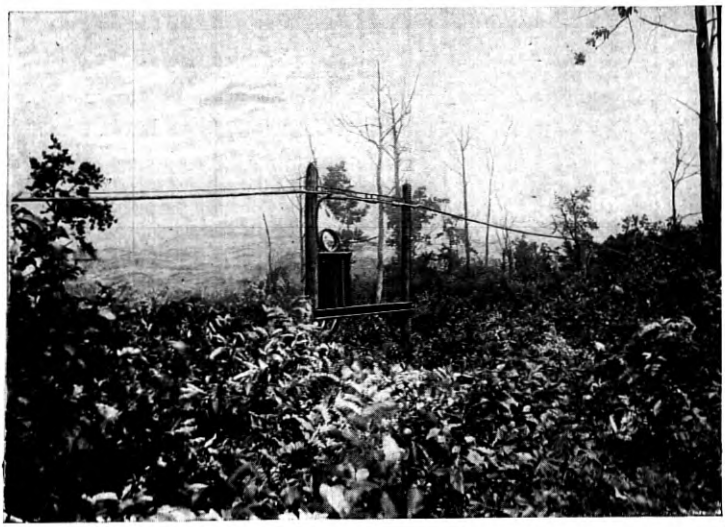


Fig. 8—Toll cable line showing loading coil case

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 1,400,000 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

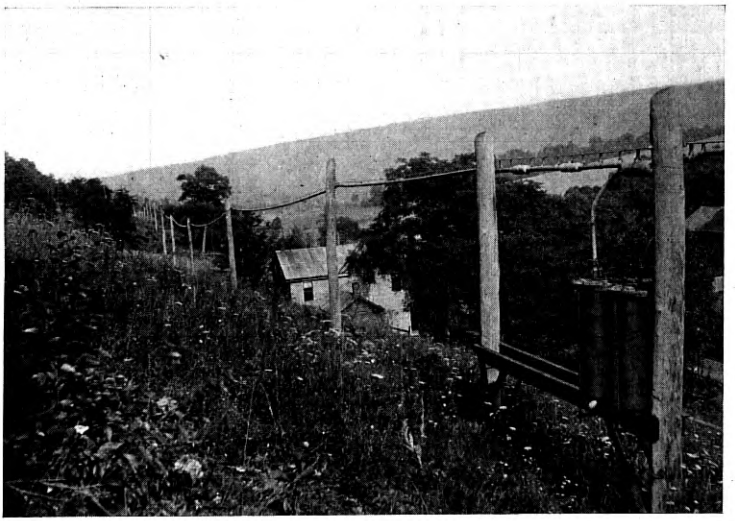


Fig. 9—Toll cable line through open country

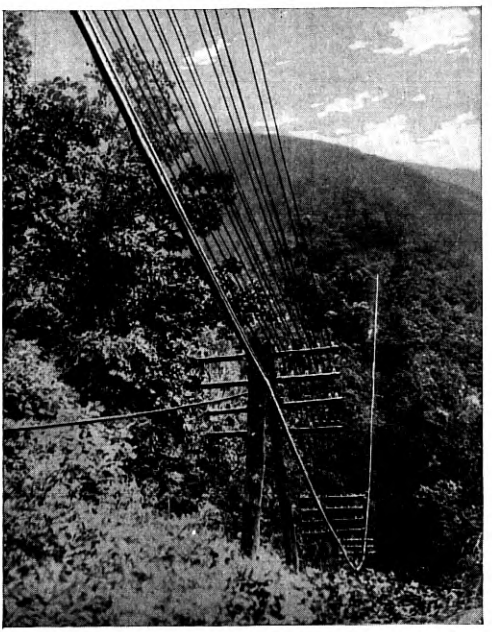


Fig. 10—Cable and open wire toll line in Allegheny Mountains

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.

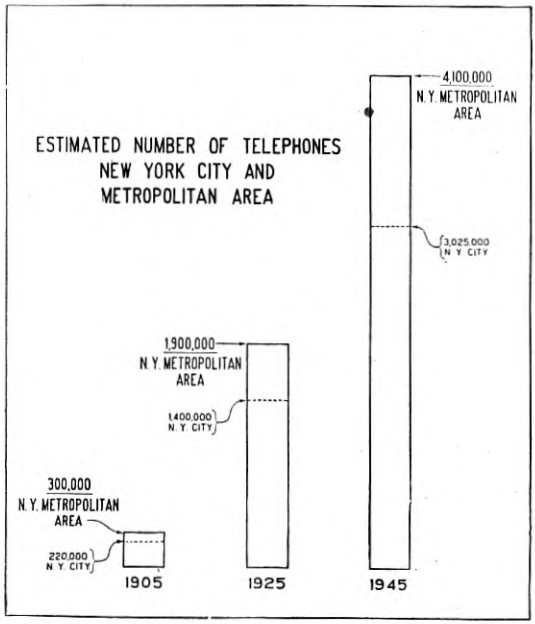


Fig. 11

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

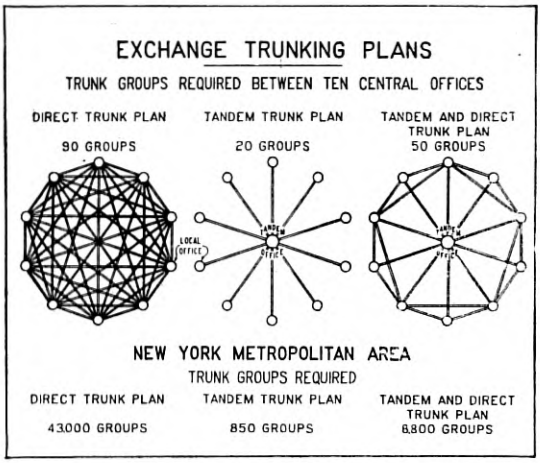


Fig. 12

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

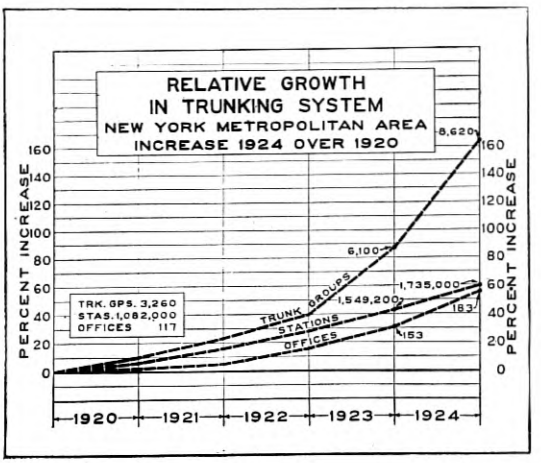


Fig. 13

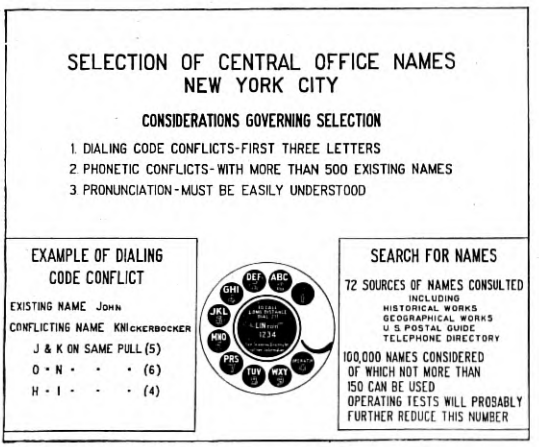


Fig. 14

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.

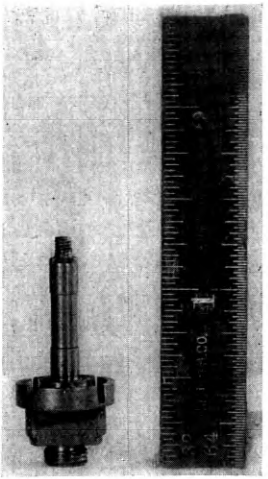
Planning for the Future

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 +.000″,-.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.*

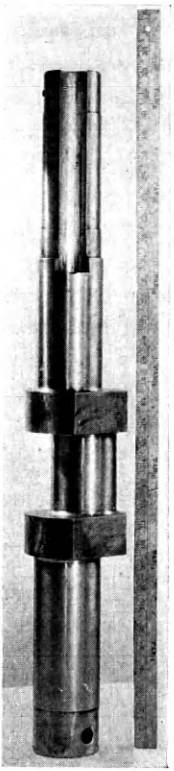
1 chuck, 2 form tools, thread die, emery stick and 4 gauges

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, ¾″ slot and 4 flats—3 oper's. Limits ±.005″. Drill ¾″ x ⅜″ 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.

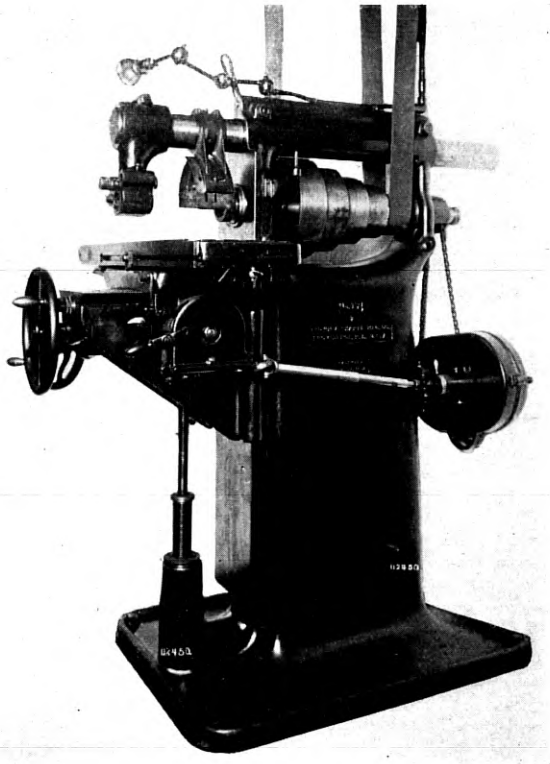


Fig. 3—Belt Driven Milling Machine

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

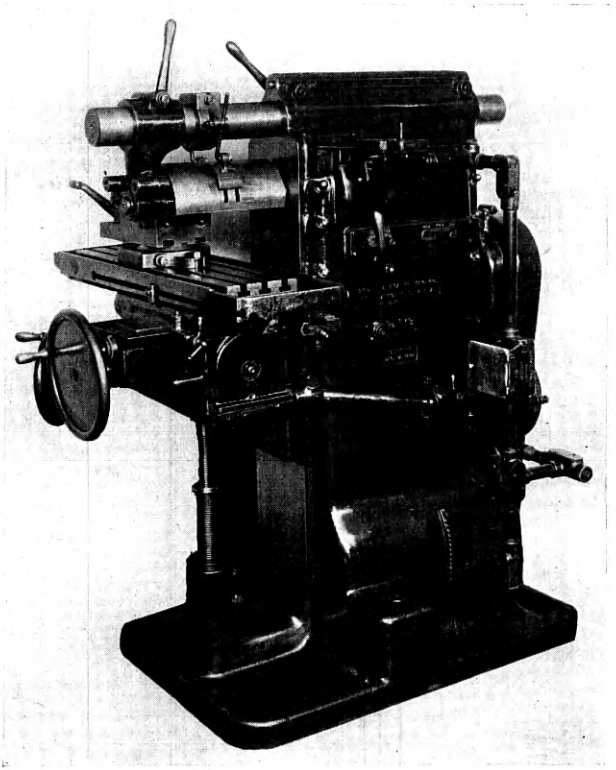


Fig. 4—Motor Driven Milling Machine

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	1/15" 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.

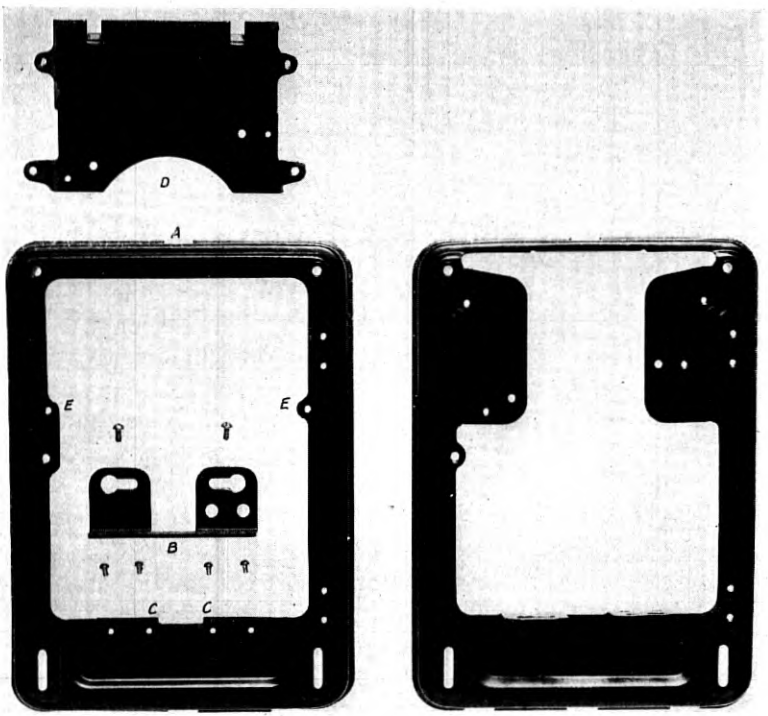
It might appear that after the careful thought already given to the methods to be employed in producing a piece of apparatus, the necessity for further study has been eliminated. This, however, is not the case, since in the original planning we must adhere closely to methods and processes that have been 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

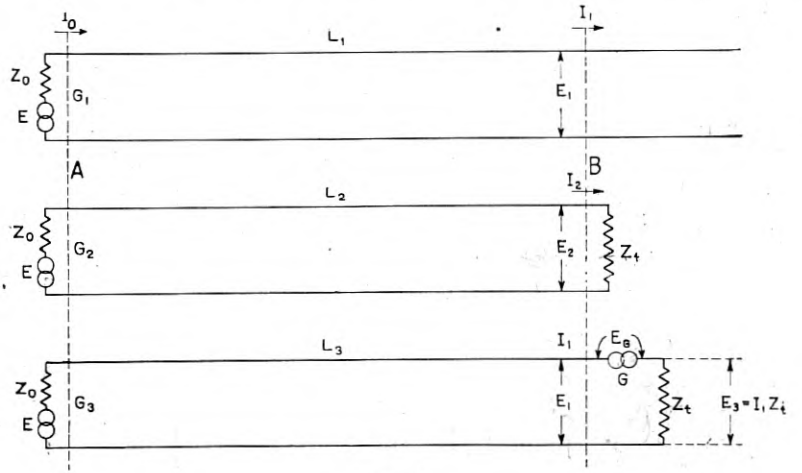


Fig. 1

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

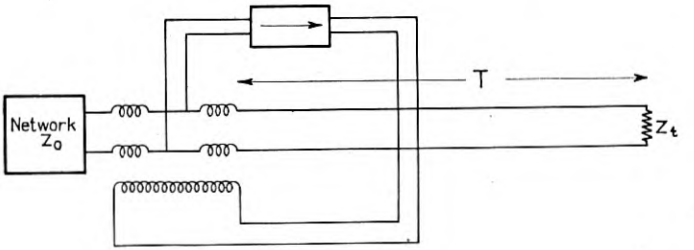


Fig. 2

the reflected and the incident current is known as the "reflection coefficient," the value of which is expressed as follows:

$$r=\frac{I_3}{I_1}=\frac{Z_o-Z_t}{Z_o+Z_t}. \tag{7}$$

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_o = 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_F = 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.

$i_1, i_2, i_3, \cdots i_n$ = Currents Returned from the 1, 2, 3, - - - - - and nth Irregularities.

$i_1', i_2', i_3', \cdots i_n'$ = In-Phase Components of $i_1, i_2, i_3, \cdots i_n$

$i_1'', i_2'', i_3'', \cdots i_n''$ = Quadrature Components of $i_1, i_2, i_3, \cdots i_n$

$k = \sqrt{\frac{L}{C}}$ = 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[2] Reflection Coefficient at Capacity Irregularities.
R_L = Representative[2] Reflection Coefficient at Inductance Irregularities.
r_C = Reflection Coefficient at a Capacity Irregularity.
r_L = Reflection Coefficient at an Inductance Irregularity.
$r_1, r_2, r_3, - - - r_n$ = 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_F = Distribution Function.
S_H = Irregularity Function.
S_w = Frequency Function.
T = Transmission Loss in a Finite Line.
$\Theta_1, \Theta_2, \Theta_3, - - - - \Theta_n$ = 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

[2] 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

$$r_C = \tfrac{1}{2} \frac{\frac{k}{\sqrt{1-w^2}}}{\frac{1}{2\pi f h_C C}} \tag{12}$$

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) representa-

tive 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

$$R_L = H_L \frac{w}{\sqrt{1-w^2}}. \tag{25}$$

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

$$I' = I_o R_L \sqrt{(\cos^2\Theta_1 + A^4\cos^2\Theta_2 + A^8\cos^2\Theta_3 + - - - - A^{4(n-1)}\cos^2\Theta_n)} \tag{26}$$

and the representative quadrature current is

$$I'' = I_o R_L \sqrt{(\sin^2\Theta_1 + A^4\sin^2\Theta_2 + A^8\sin^2\Theta_3 + - - - - A^{4(n-1)}\sin^2\Theta_n)}. \tag{27}$$

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,

$$I' \neq I'' = \sqrt{\frac{I'^2 + I''^2}{2}} = \frac{I_o R_L}{\sqrt{2}} \sqrt{1 + A^4 + A^8 + - - - - - - - A^{4(n-1)}} \tag{28}$$

For a finite number of irregularities, that is a finite line terminated by a perfect network just beyond the nth coil:

$$I' = I'' = \frac{I_o R_L}{\sqrt{2}} \sqrt{\frac{1-A^{4n}}{1-A^4}} \tag{29}$$

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

$$I'_\infty = I''_\infty = \frac{I_O R_L}{\sqrt{2}} \sqrt{\frac{1}{1-A^4}}. \tag{30}$$

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

$$S=10\log_{10}\left[\frac{1_o^2}{2I'^2\log_e\frac{1}{F}}\right]. \tag{39}$$

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 40 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

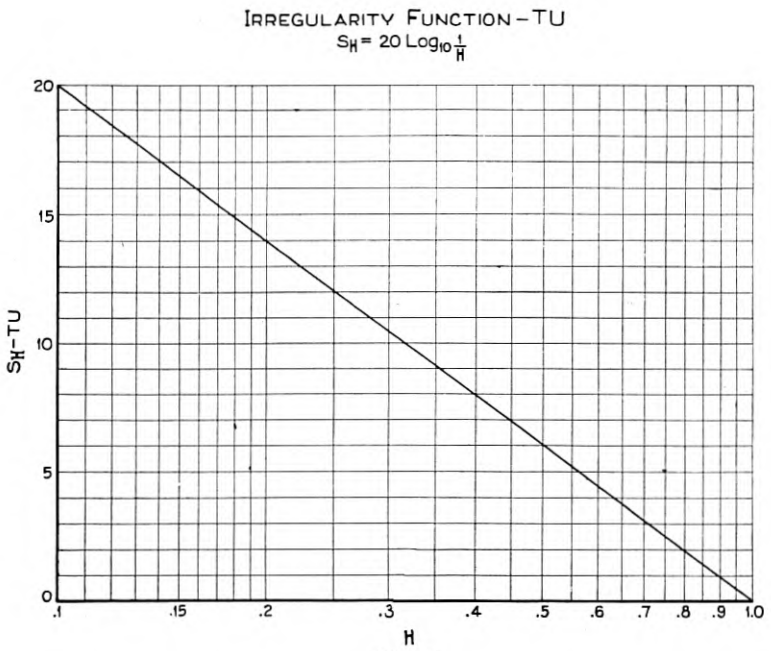

IRREGULARITY FUNCTION – TU
$S_H = 20 \log_{10} \frac{1}{H}$
S_H-TU
20
15
10
5
0
.1
.15
.2
.3
.4
.5
.6
.7
.8
.9
1.0
H

Fig. 3

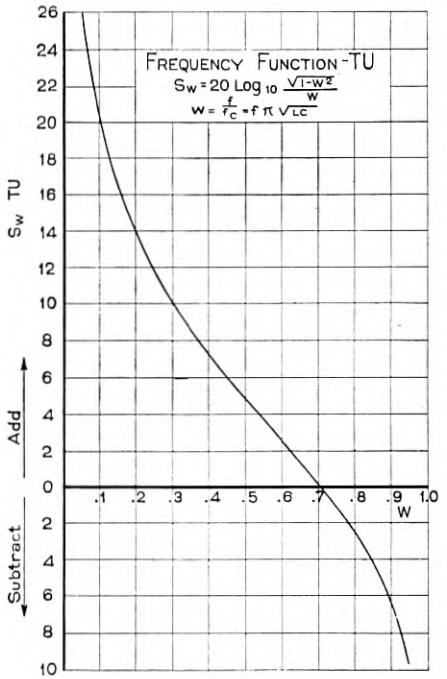

FREQUENCY FUNCTION -TU
$S_W = 20 \log_{10} \frac{\sqrt{1-W^2}}{W}$
$W = \frac{f}{f_c} = f \pi \sqrt{LC}$
S_W TU
Add
Subtract
26
24
22
20
18
16
14
12
10
8
6
4
2
0
2
4
6
8
10
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0
W

Fig. 4

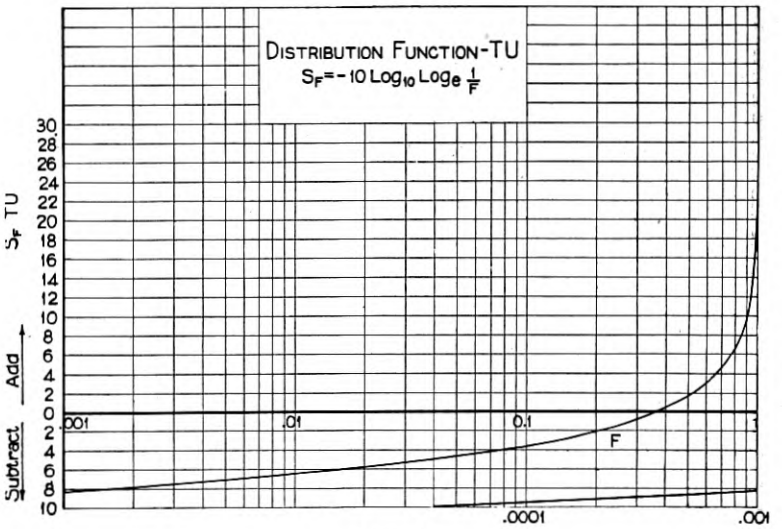


Fig. 5

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,

$L = 20 \log_{10} \frac{1}{A}$ = loss per loading section in TU

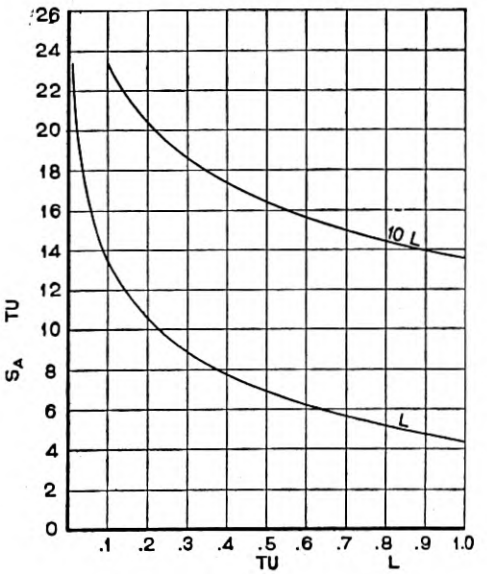


Fig. 6

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.

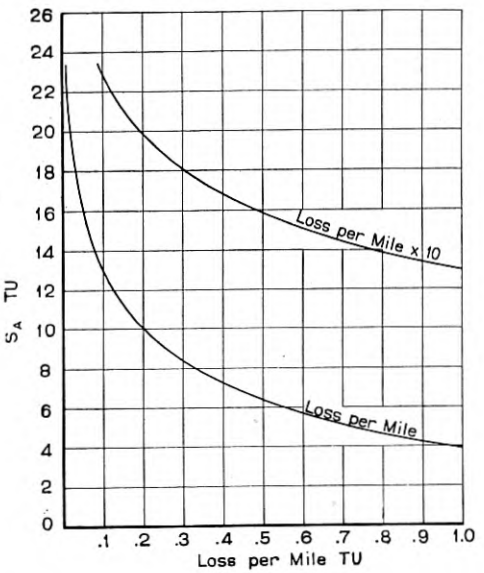


Fig. 7

Increase of the return loss when the line is limited to n sections

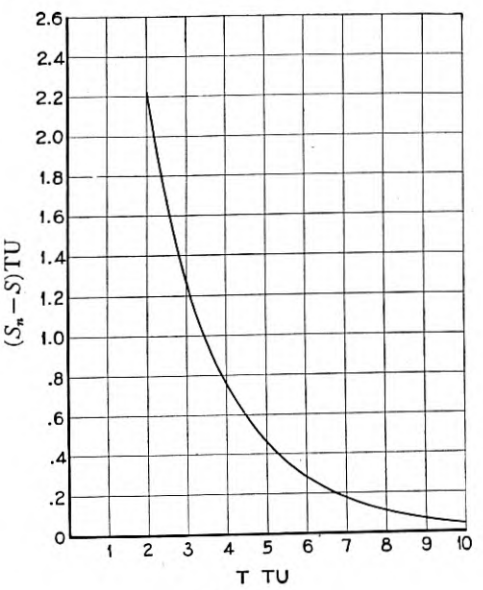


Fig. 8

(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 $w = \frac{1000}{2810} = 0.356.$

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 0.274. 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, $0.274 \times \frac{6000}{5280} = 0.311$ 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 (42) 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 14, 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

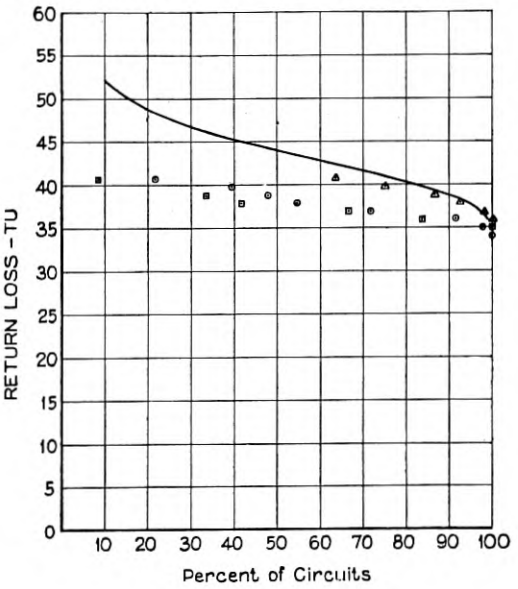


Fig. 9

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

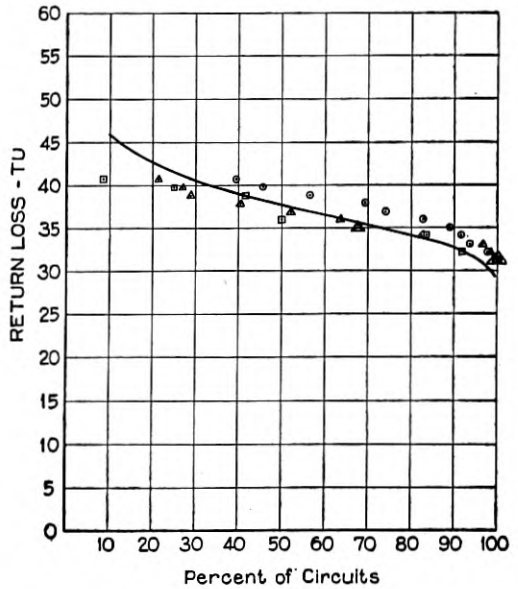


Fig. 10

Return loss of No. 19–H–174–63 sides exceeded by various percentages of circuits at 2000 cycles

Smooth curve—theoretical

⊙ 46-H-174–63 sides Pittsburgh to Ligonier

⊡ 12–H–174–63 sides Ligonier to Pittsburgh

△ 52–H–174–106 sides Pittsburgh to Ligonier

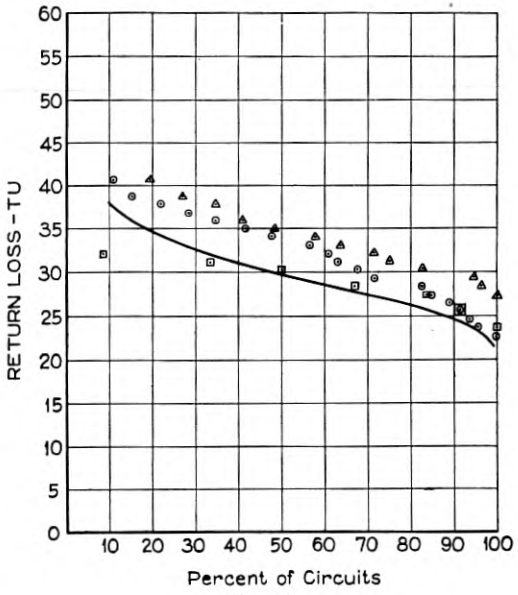


Fig. 11

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

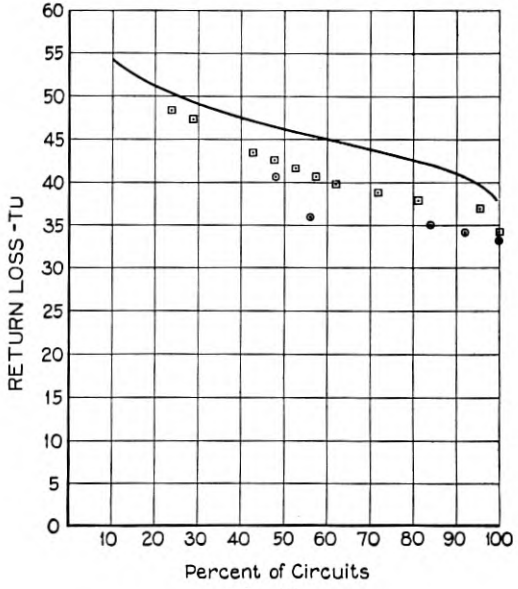


Fig. 12

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

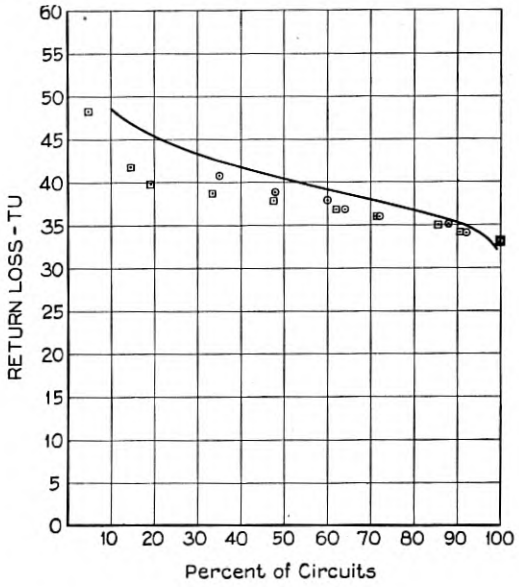


Fig. 13

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

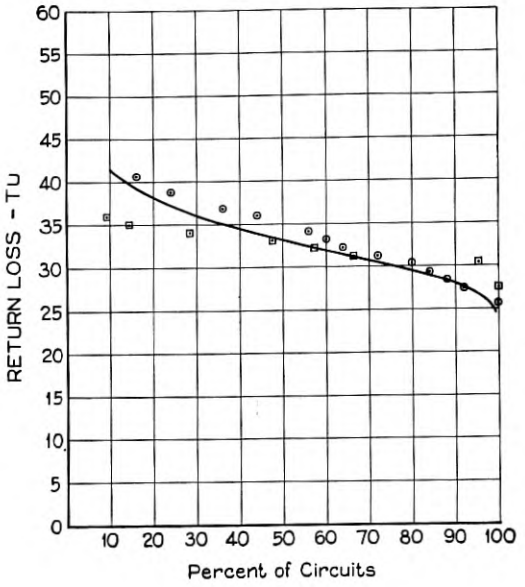


Fig. 14

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."

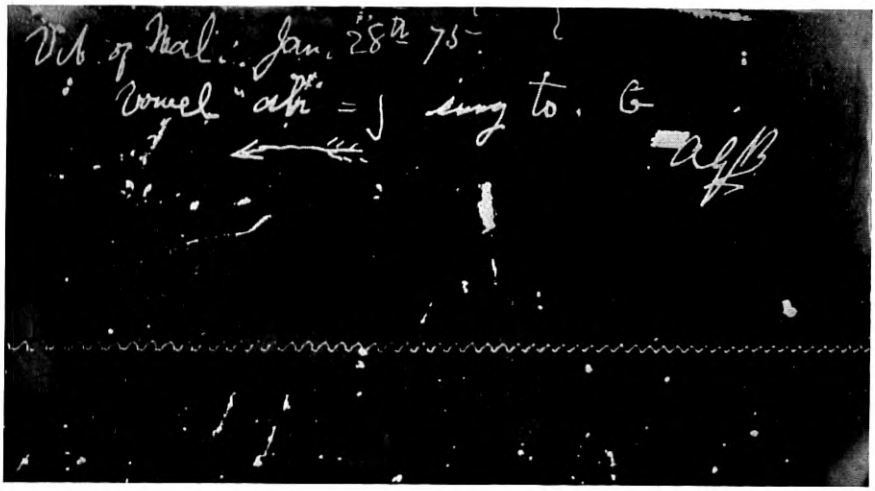
In the accompanying paper, Dr. I. B. Crandall describes modern methods whereby with the most refined apparatus, highly accurate speech wave forms have been produced. The analysis and interpretation of both vowel and consonant sounds made possible by these records, are the realization of an objective 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

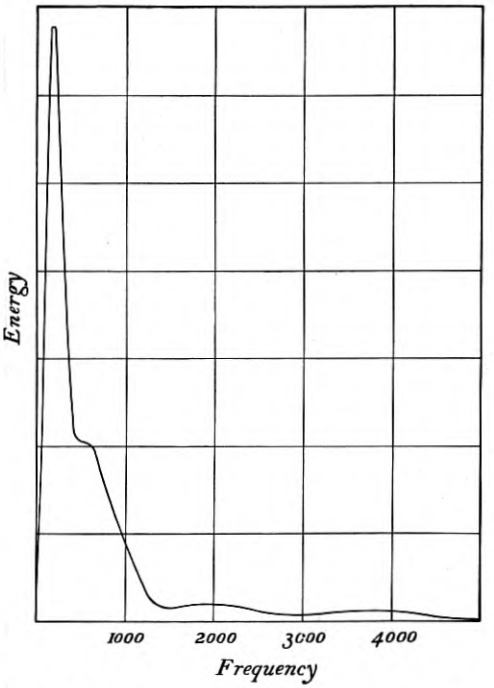


Fig. 1—Energy distribution; composite curve for male and female voices

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. 8i) 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 oscilla-

tory 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,

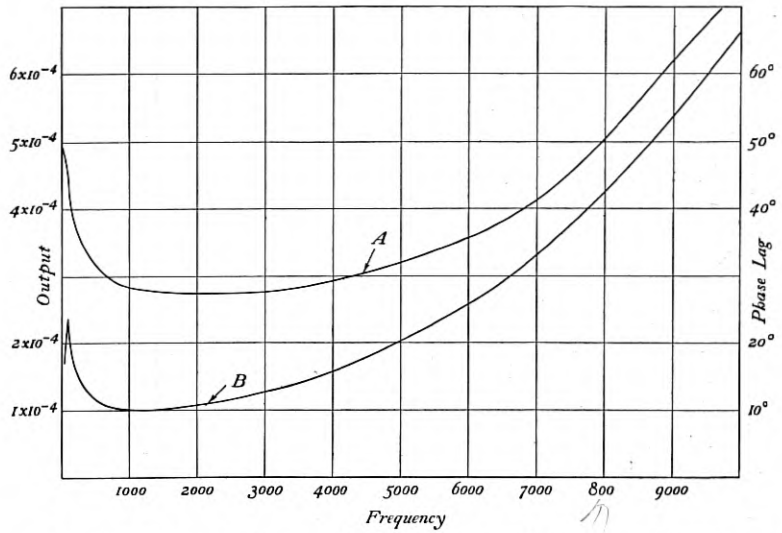


Fig. 2—Curve *A*: Output of transmitter in volts per dyne per sq. cm. Curve *B*: Phase lag of voltage behind pressure in condenser transmitter

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

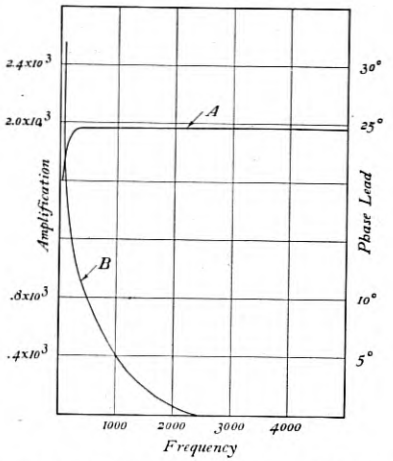


Fig. 3—Curve *A:* Amplitude frequency characteristic of amplifier. Curve *B:* Phase lead of output, vs. frequency of voltage input to amplifier

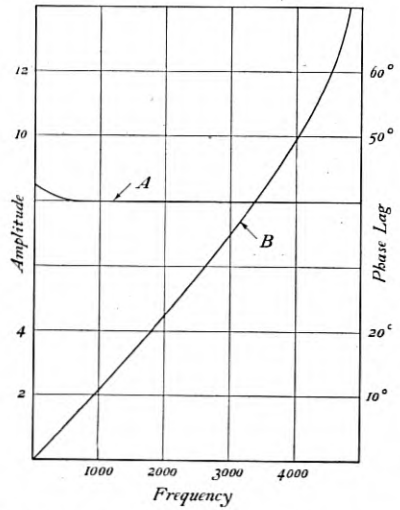


Fig. 4—Curve *A:* Amplitude frequency characteristic of oscillograph. Curve *B:* Phase lag of amplitude behind current in oscillograph

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

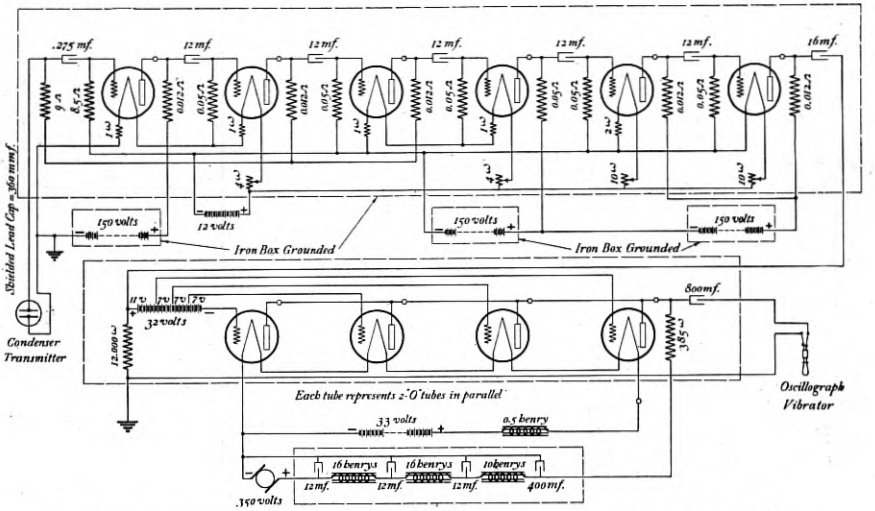


Fig. 5—General diagram of recording apparatus showing circuit details

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. 4. 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

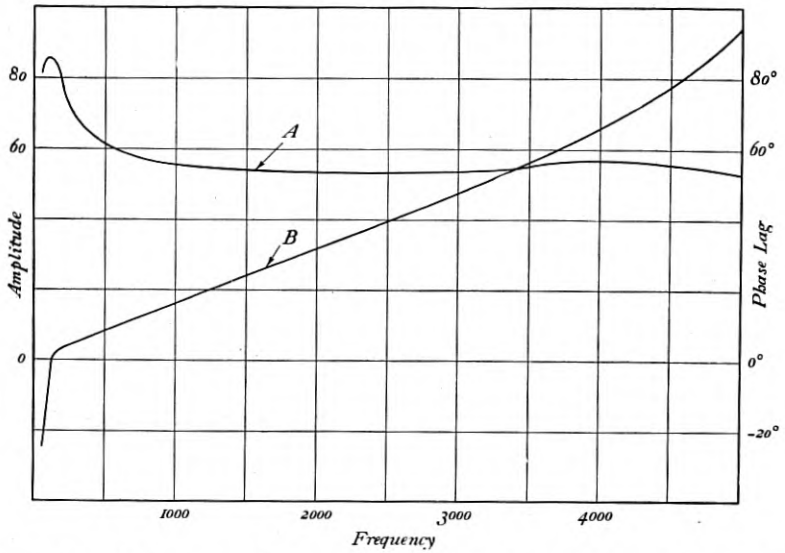


Fig. 6—Overall frequency characteristics of amplitude and phase of the recording system. Curve *A:* Oscillographic amplitude per unit of pressure on transmitter diaphragm. Curve *B:* Phase lag of oscillographic amplitude behind pressure on diaphragm

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

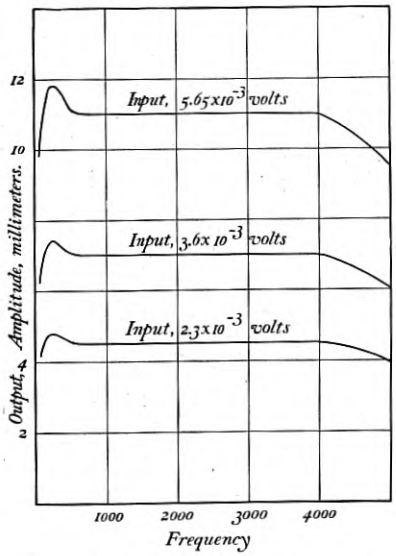


Fig. 7—Amplitude frequency characteristics of circuit-oscillograph at different energy levels

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.

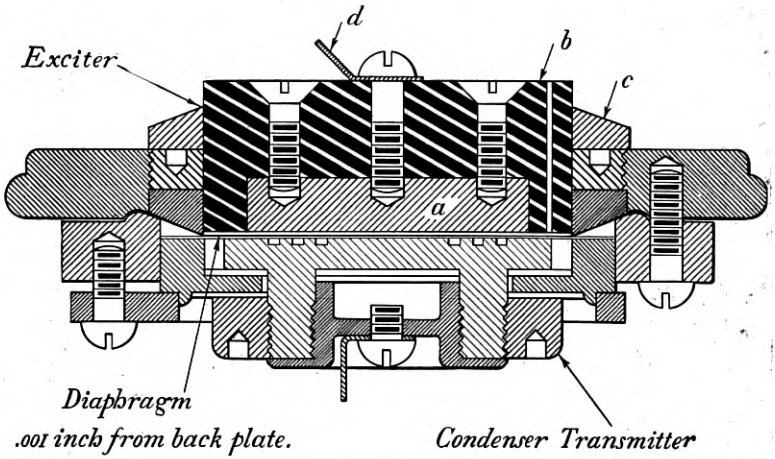


Fig. 8—Condenser transmitter coupled with square-topped-wave exciter

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

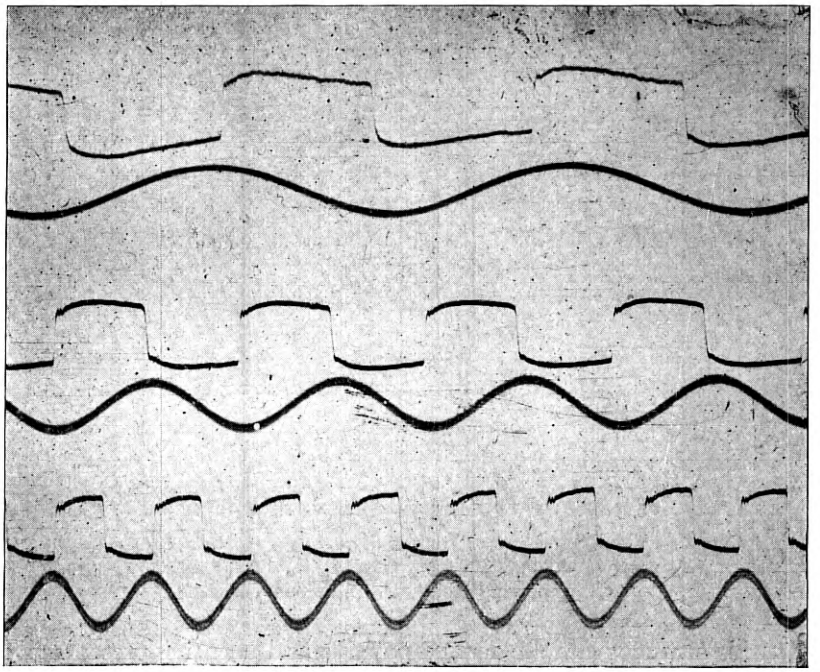


Fig. 9—Oscillogram of square-topped acoustic waves as recorded by the apparatus

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

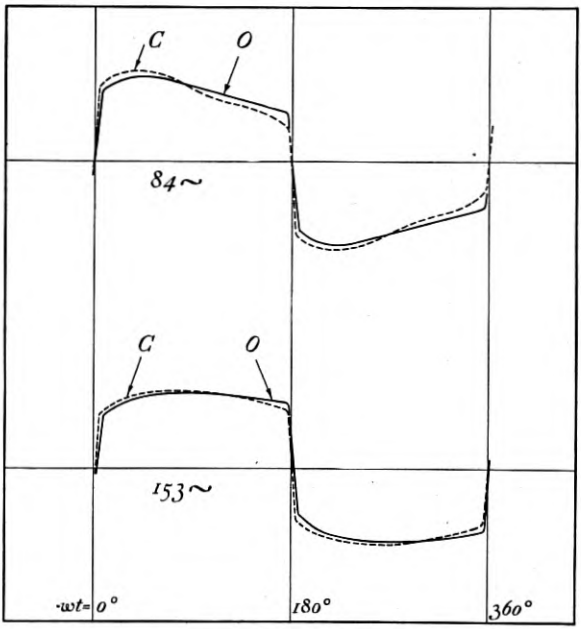


Fig. 10—Calculated and observed wave forms, as recorded by the apparatus

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

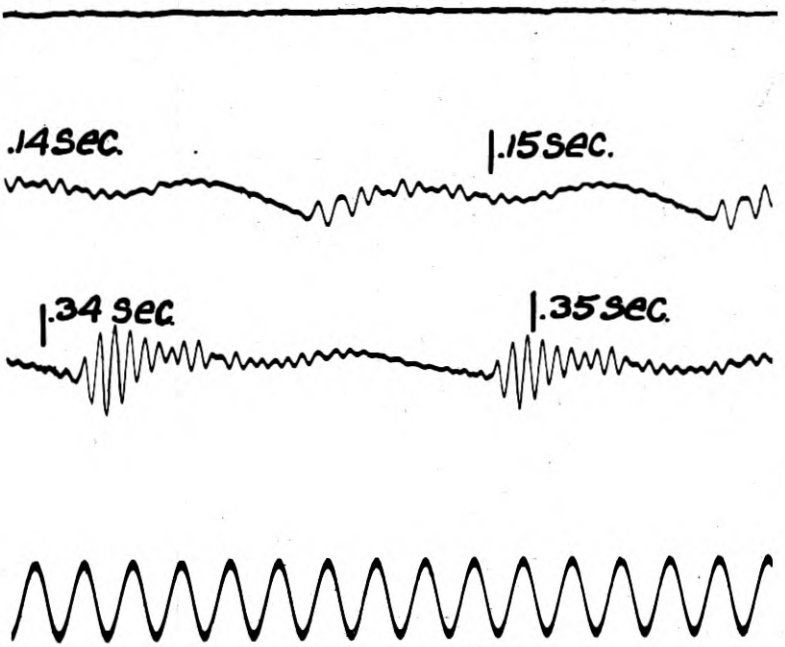


Fig. 11—Section of original record showing timing wave

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

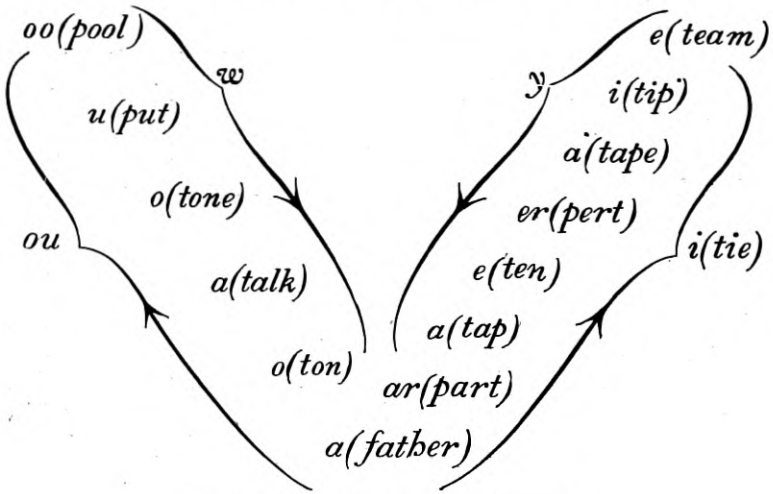


Fig. 12—Classification of vowel sounds

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

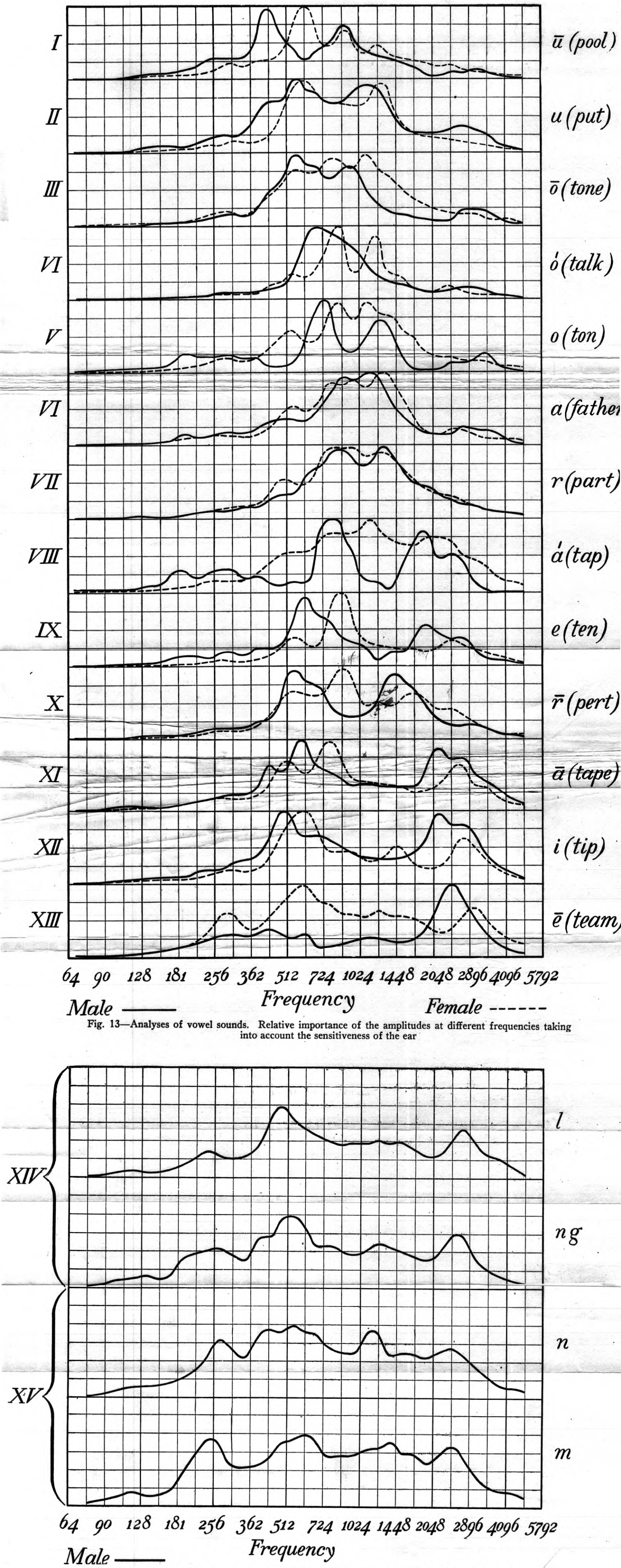


Fig. 13—Analyses of vowel sounds. Relative importance of the amplitudes at different frequencies taking into account the sensitiveness of the ear

Fig. 14—Analysis of four semi-vowel sounds. Relative importance of the amplitudes at different frequencies taking into account the sensitiveness of the ear

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	40	*o* as in *ton*	FD
139	41	*a* as in *father*	MA
151	49	*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 49) which was chosen is the most characteristic and interesting one of its group. The other vowel records (Plates 9, 40, 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 40) 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 104 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 104 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 per-

formance 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. 24, 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 0.04 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*(p*oo*l)	.061	.164	.126	.351	140	270	411	581	750 (1)	1200 (1)			3700 (4)	4412 (4)
II *u*(p*u*t)	.057	.115	.077	.249	138	250	457	691	988 (4)	1100 (3)			3637 (4)	4250 (4)
III *o*(t*o*ne)	.053	.139	.133	.325	116	237	520	729	830 (3)	1112 (4)			3475 (4)	3700 (4)
IV *a*(t*a*lk)	.034	.191	.065	.290	112	243	722	801	950 (2)	1150 (2)			3612 (4)	4075 (4)
V *o*(t*o*n)	.046	.179	.061	.280	118	253	654	854	1100 (4)	1188 (4)			3212 (4)	3353 (3)
VI *a*(f*a*ther)	.029	.199	.078	.306	113	234	955	1036	1150 (2)	1425 (2)			3683 (4)	4200 (3)
VII *ar*(p*ar*t)				.345	110	231	630	701	917 (3)	1012 (4)	1965	2162	3800 (2)	4150 (1)
									Note 1		Note 1			
VIII *a*(t*a*p)	.038	.180	.076	.294	123	232	796	960			1900	2165	3150 (3)	3175 (2)
IX *e*(t*e*n)	.034	.119	.066	.219	121	247	612	775			1800	2000	2925 (4)	2925 (4)
											Note 2		Note 2	
X *er*(p*er*t)				.331	131	239	570	712			1688	2188	3050 (2)	3500 (1)
XI *a*(t*a*pe)	.042	.172	.091	.305	125	235	494	614			3000	2800		
XII *i*(t*i*p)	.036	.126	.049	.211	137	253	450	523			2950	2962		
XIII *e*(t*ea*m)	.036	.189	.116	.341	136	252	296	332			2987	3266		4800 (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	*1688–1965*	*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 (§I) 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 *ð* [*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 4600 cycles the outstanding phenomena of *zh/sh* are well defined. In addition to frequencies of 2048-3249 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 4200-8000 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.

TABLE VI

Group XVI—6 Stop Consonants; Transitional dth/th

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 6)	No. of Cycles	First Cycle Short	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	92,184	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	104	112
136	*ta*	MB	.06	unvoiced	none	unvoiced	3600	900	3000	2	yes	120	113
137	*ga*	MA	.12	100,200,300	none	84,252	1600, 2800 (Note 4)	550	3000	3	...	101	111
138	*ga*	MB	.10	100,200,300	none	95,190	1400, 4000	600	3600	2	...	112	112
139	*ka*	MA	.07	unvoiced	none	unvoiced	1500, 4000 (Note 5)	1200	3800	4	yes	109	118
140	*ka*	MB	.08	unvoiced	none	unvoiced	1600, 4200	1300	4000	4	yes	125	116
141	*dtha*	MA	.20	83,166	4000 (Note 1)	95,189	4200 (Note 1)	600	3000	2	...	104	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	(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	*va*	MA	.20	97,195,390	3000	87,174	none	600	2700	3	...	101	116
146	*va*	MB	.25	112,224	3200 (trace)	100,200	none	600	3400	2	...	112	107
147	*fa*	MA	.15	unvoiced	3100	unvoiced	3500, 7000	500	2800	4	yes	112	121
148	*fa*	MB	.30	irregular	3200, 6400	unvoiced	3200, 6400	600	3600	3	yes	111	104
149	*ja*	MA	.22	81,243	3400	81,162	2600, 5200	450	2700	4	...	110	110
150	*ja*	MB	.14	trace	3300	90,179	2000, 4800	500	3100	4	...	115	111
151	*cha*	MA	.07	unvoiced	4800	unvoiced	2800, 4800	500 1500	3000	2	yes	104	111
152	*cha*	MB	.08	unvoiced	3600	unvoiced	3600, 6400	500 1600	trace	2	yes	119	115
153	*zha*	MA	.28	86,172,344	3000, 4000 (Note 1)	87	3000, 4000 (Note 1)	450	2900	4	...	100	111
154	*zha*	MB	.13	96	2600, 4200	99	3000, 4200	500 2000	....	4	...	114	111
155	*sha*	MA	.18	unvoiced	2800, 3600 (Note 2)	unvoiced	2800, 4600 (Note 2)	450	3200	3	yes	104	104
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)	400	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	114
160	*sa*	MB	.19	unvoiced	4000, 6400	unvoiced	4200, 6600	650	2900	2	yes	117	108

NOTE 1—Alternating; lower frequency in first part of fundamental cycle, higher frequency in latter part of cycle.
NOTE 2—Alternating, irregularly.
NOTE 3—Possibly due to the *a* sound.

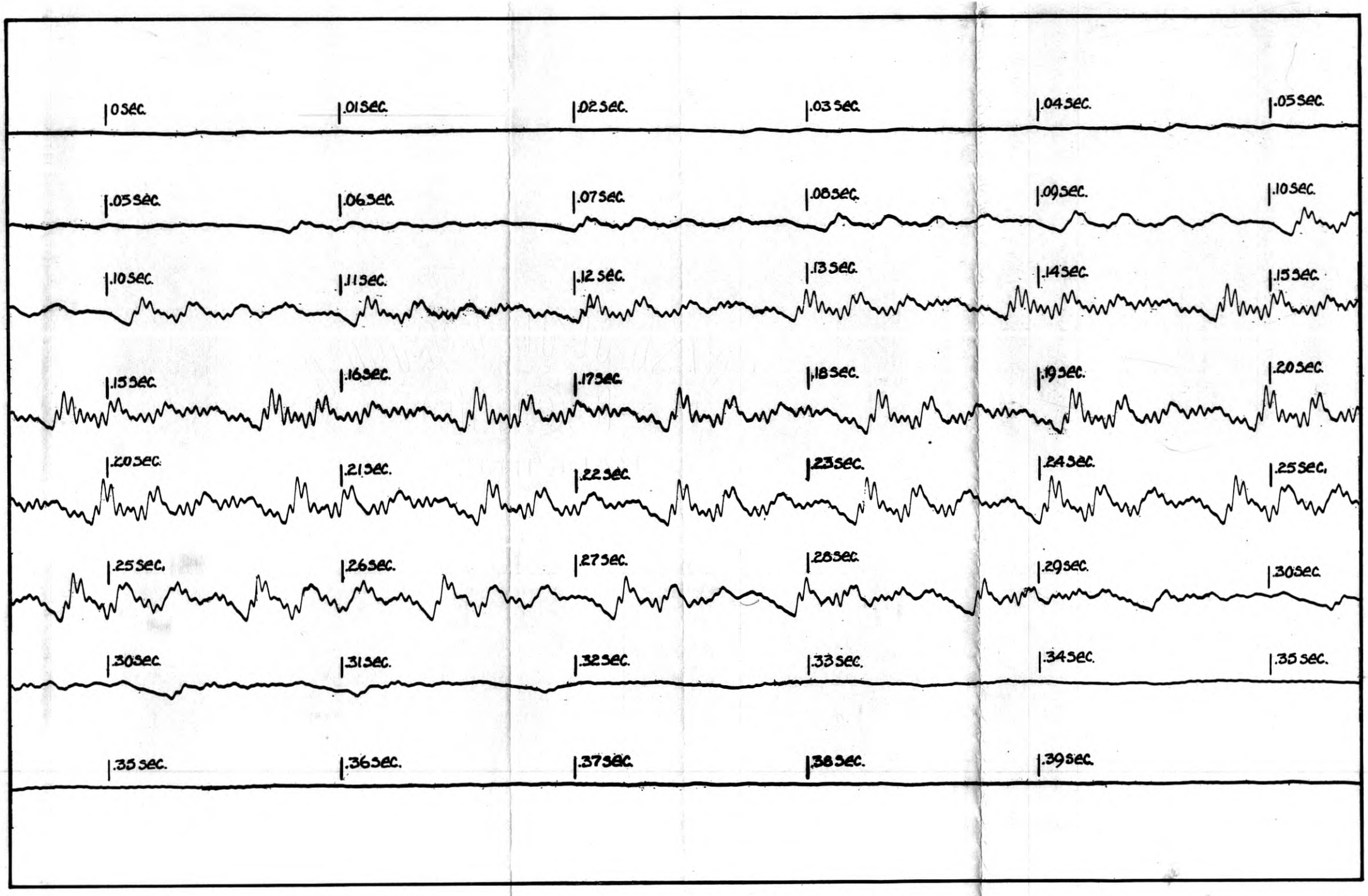


Plate No. 9—*u* as in *put*. Spoken by M.A.–Male, low-pitched

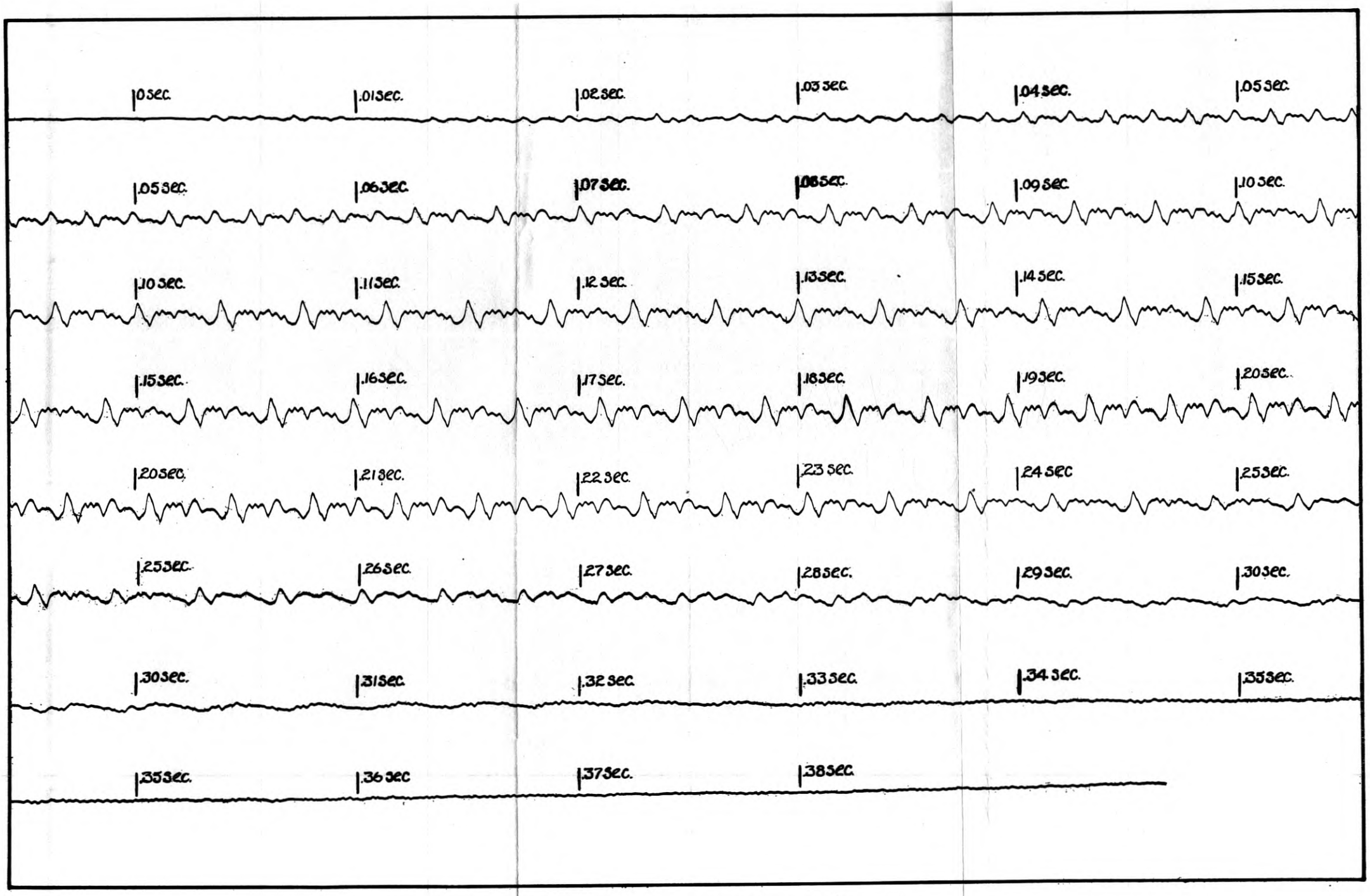


Plate No. 40—*o* as in *ton*. Spoken by F.D.–Female, high-pitched

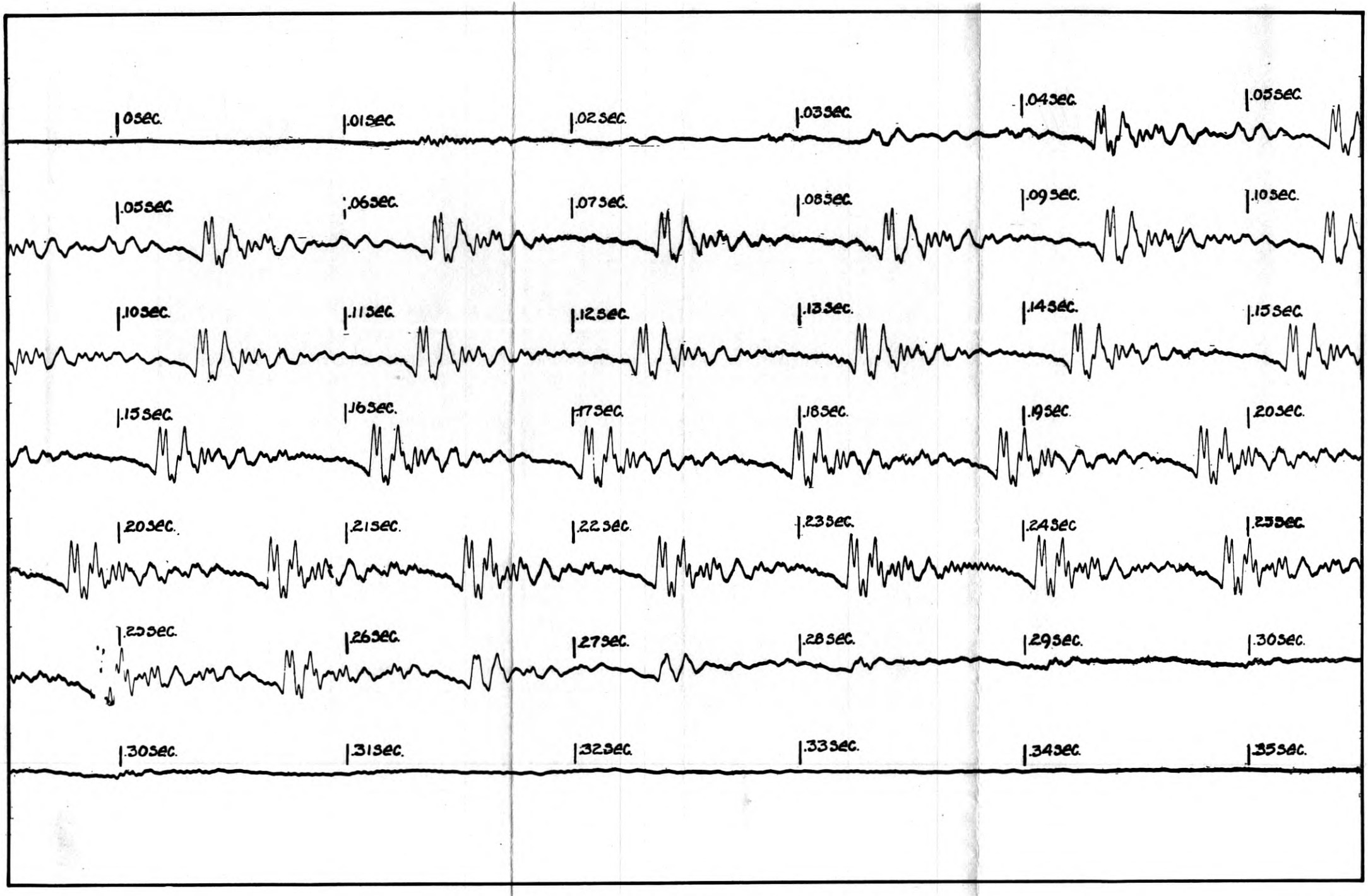


Plate No. 41—*a* as in *father*. Spoken by M.A.–Male, low-pitched

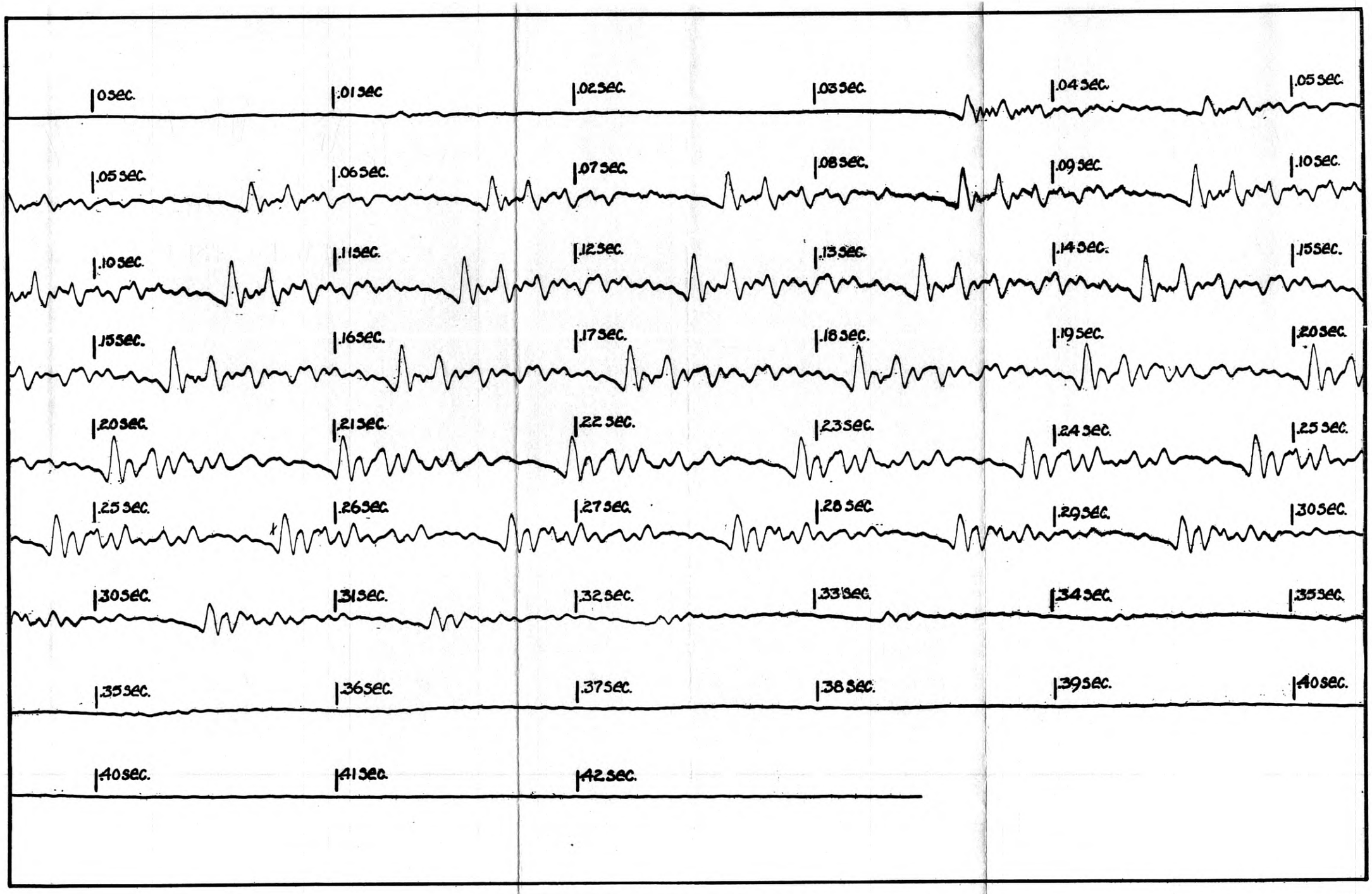


Plate No. 49—*ar* as in *part*. Spoken by M.A.–Male, low-pitched

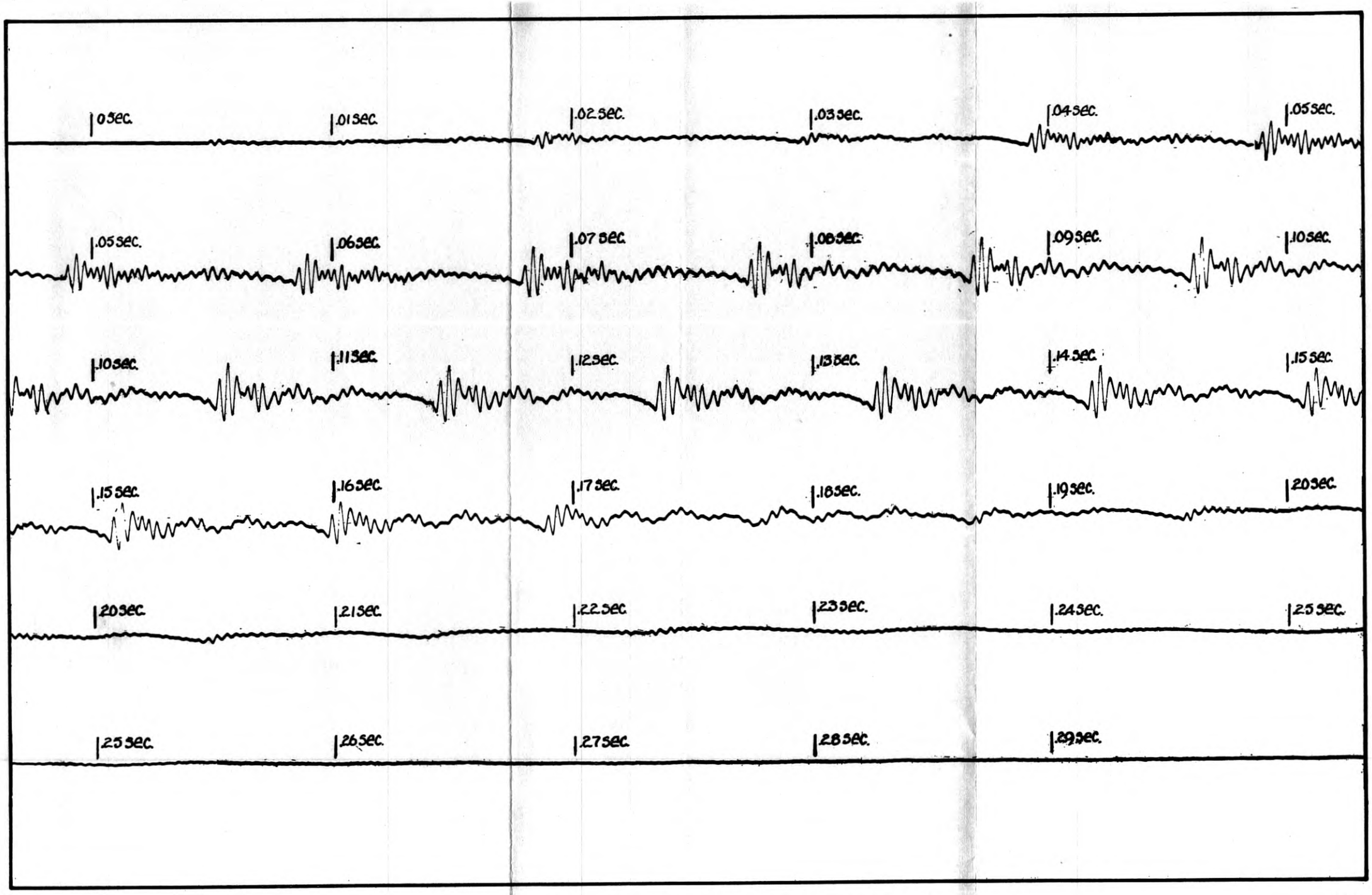


Plate No. 89—*i* as in *tip*. Spoken by M.A.–Male, low-pitched

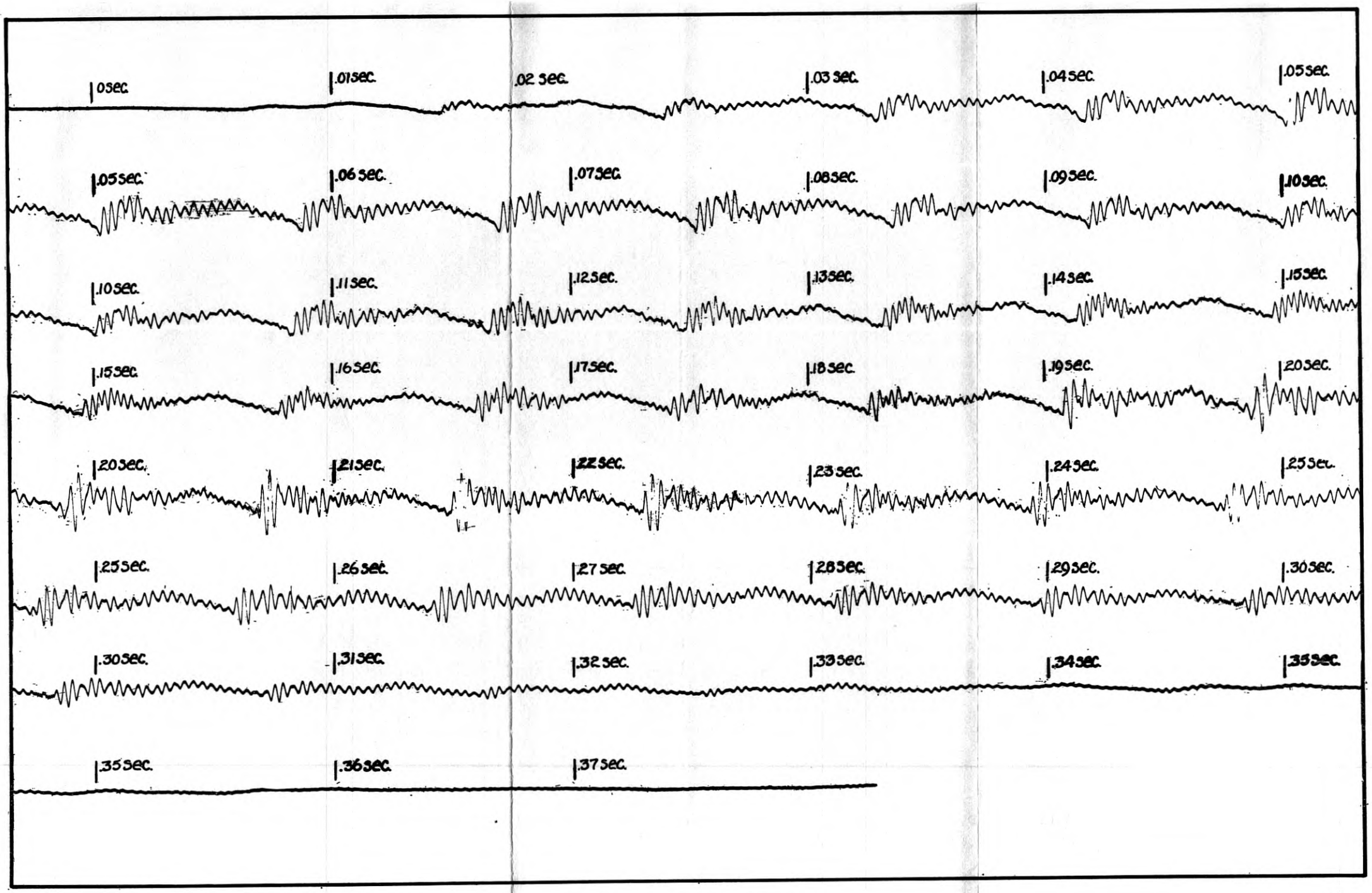


Plate No. 108—*lee*. Spoken by M.B.

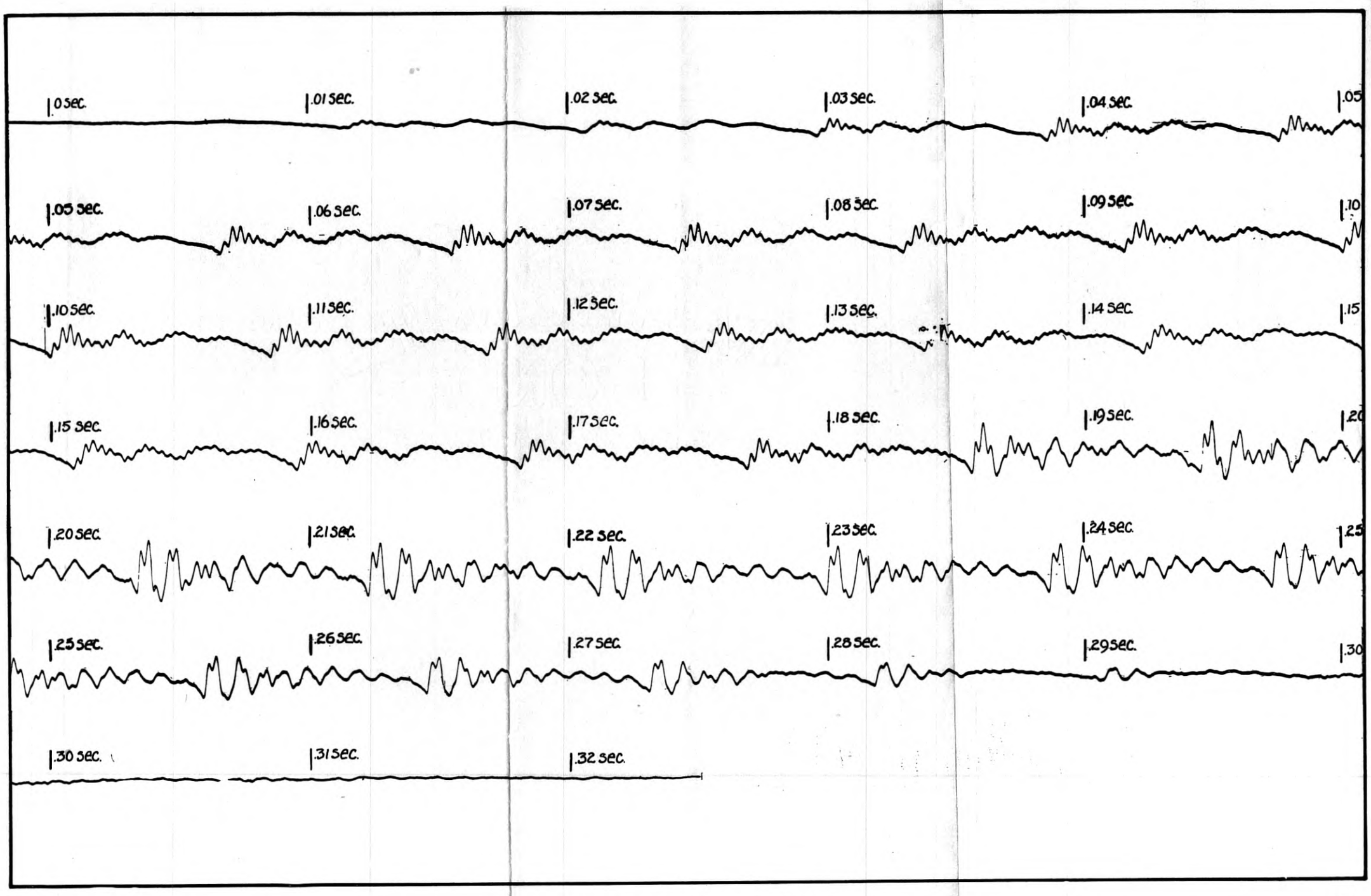


Plate No. 110—*la*. Spoken by M.B.

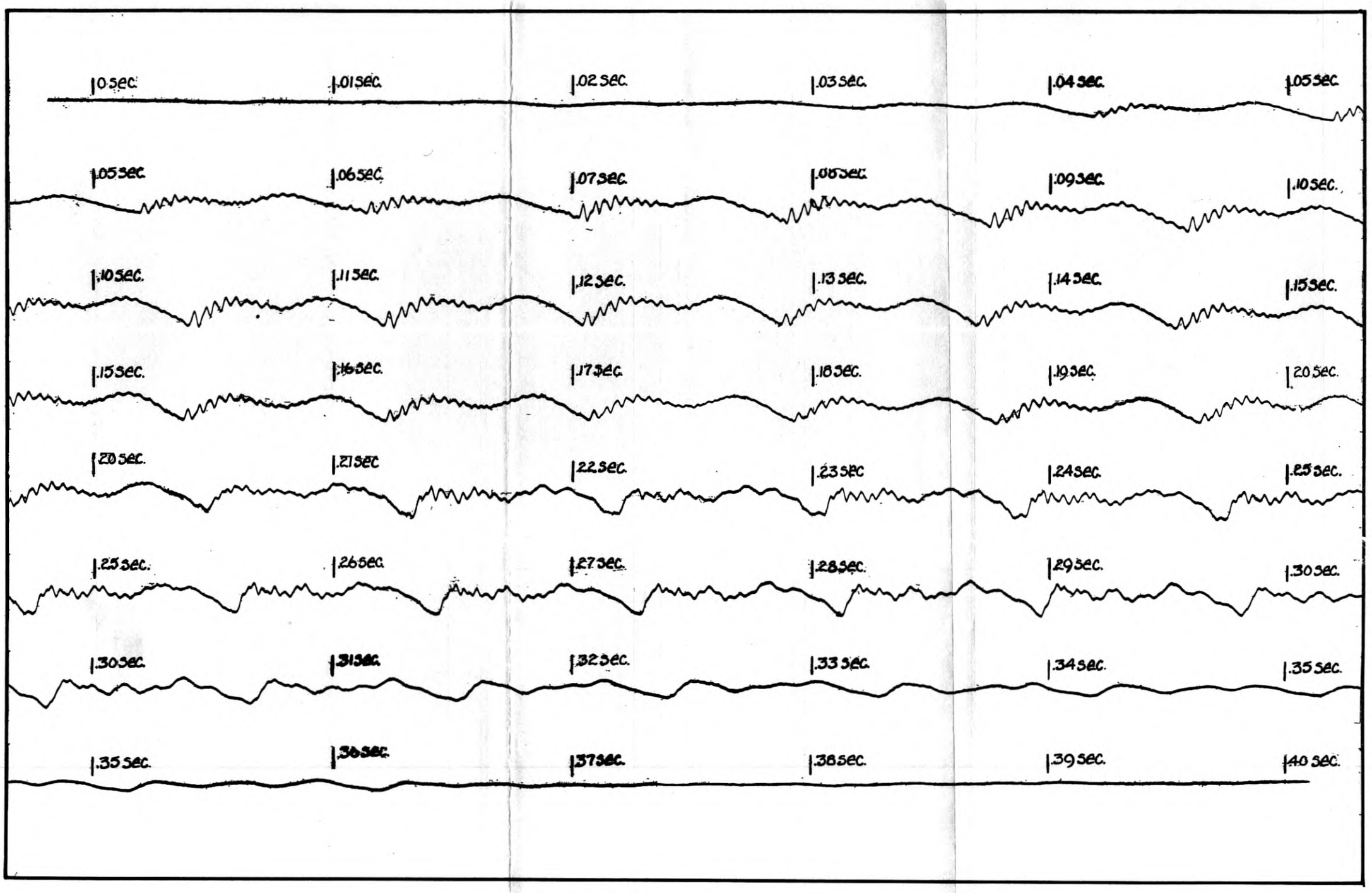


Plate No. 124—*moo*. Spoken by M.B.

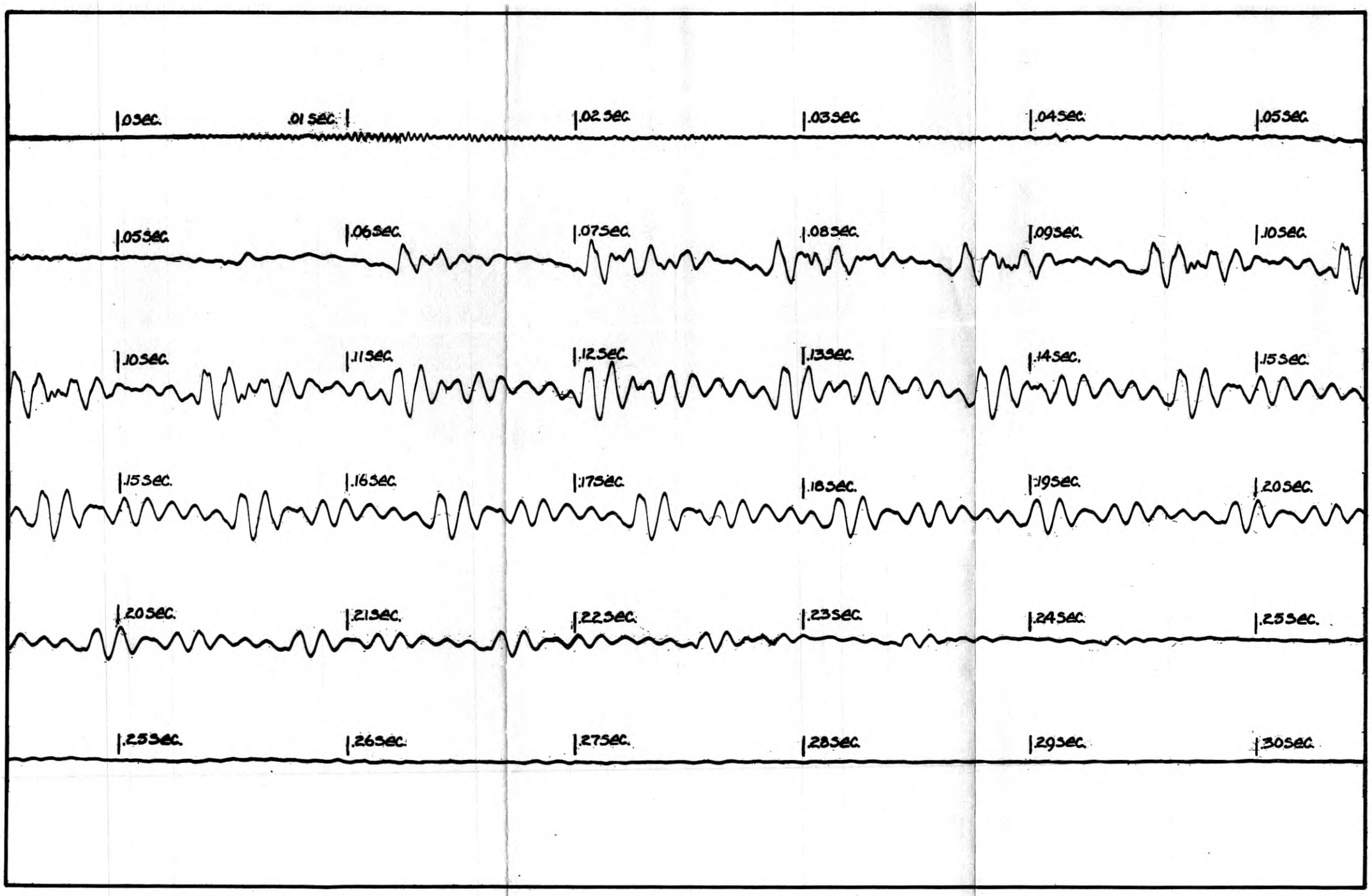


Plate No. 136—*ta*. Spoken by M.B.

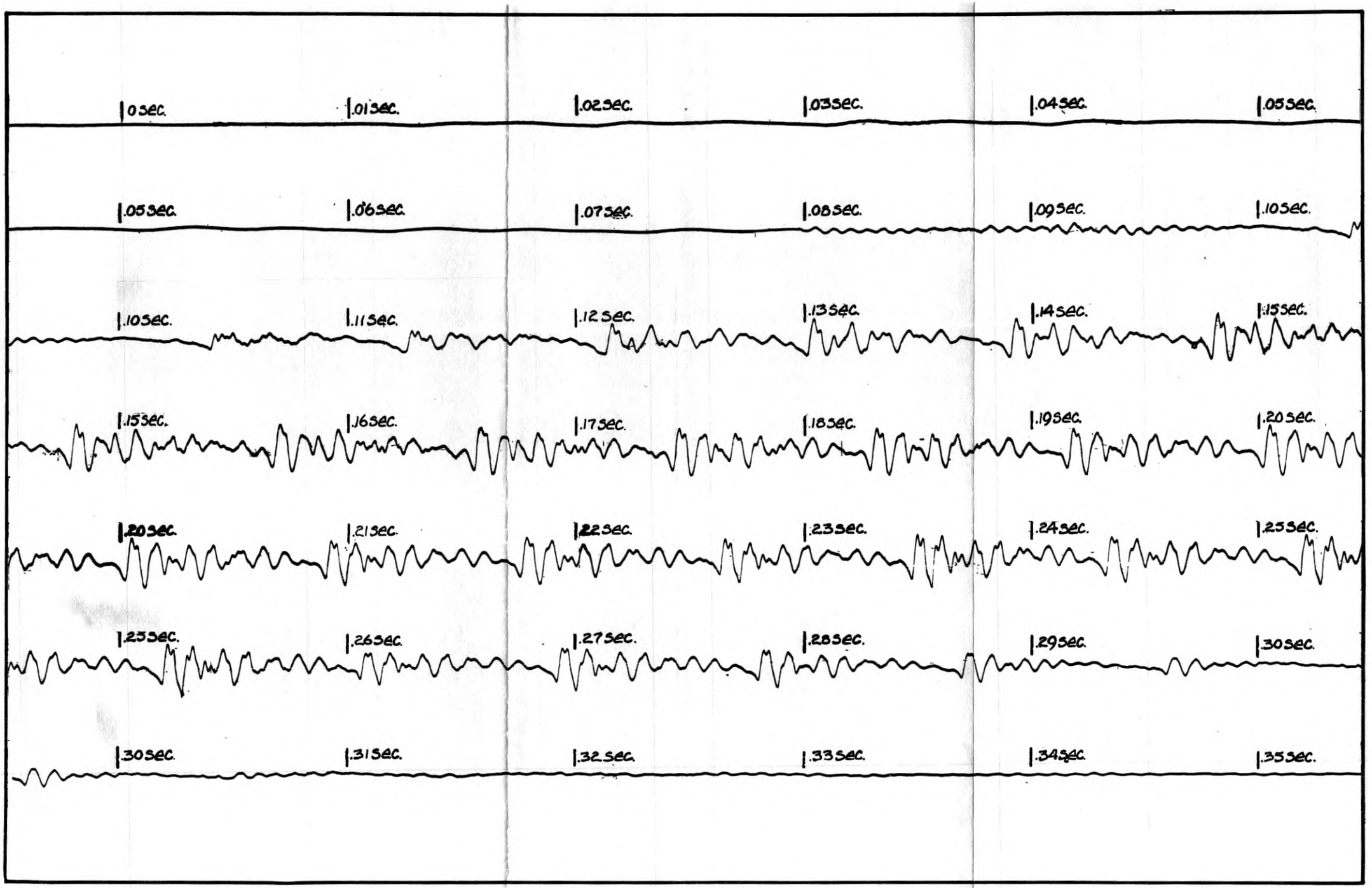


Plate No. 138—*ga*. Spoken by M.B.

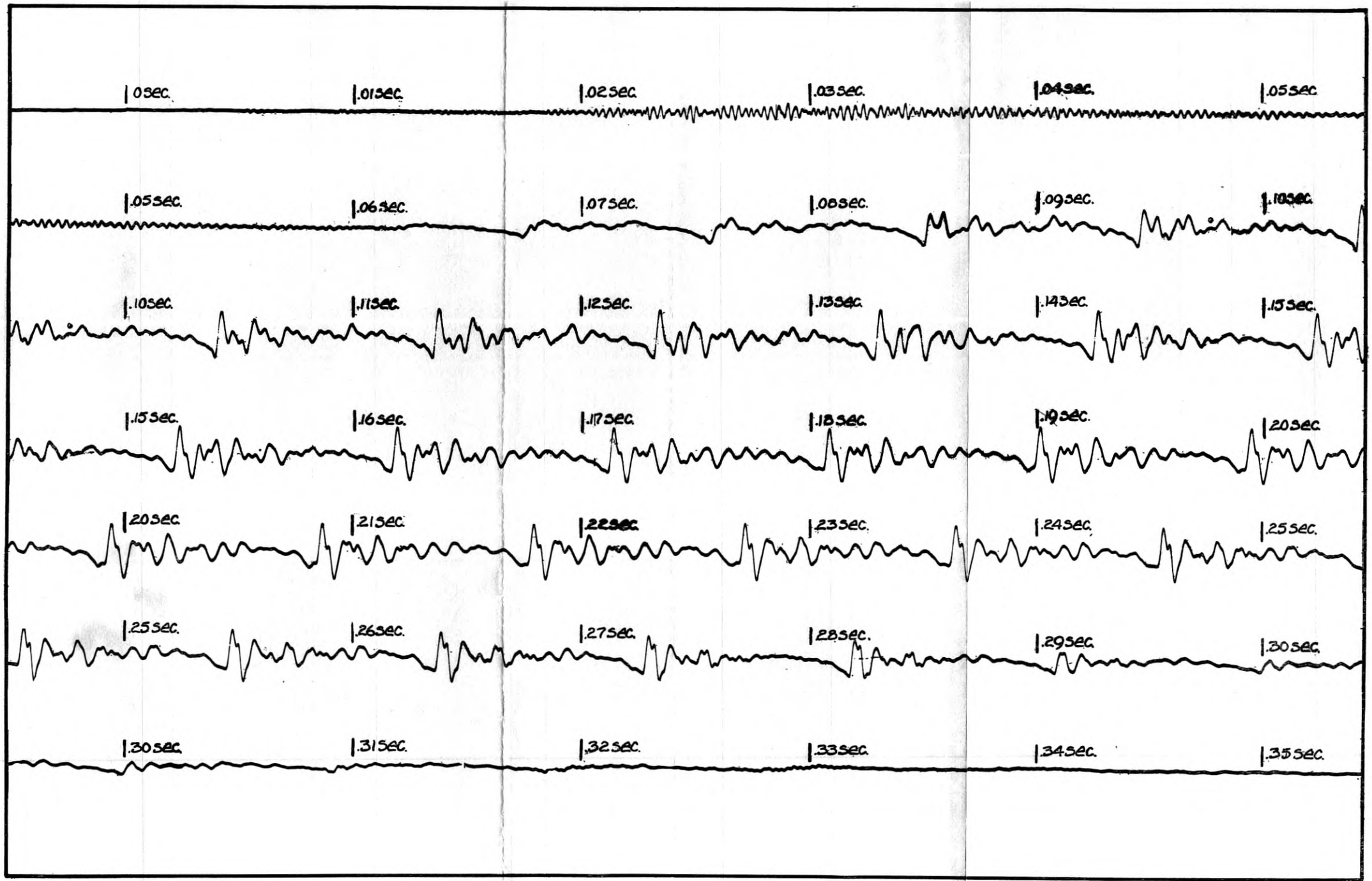


Plate No. 151—*cha.* Spoken by M.A.

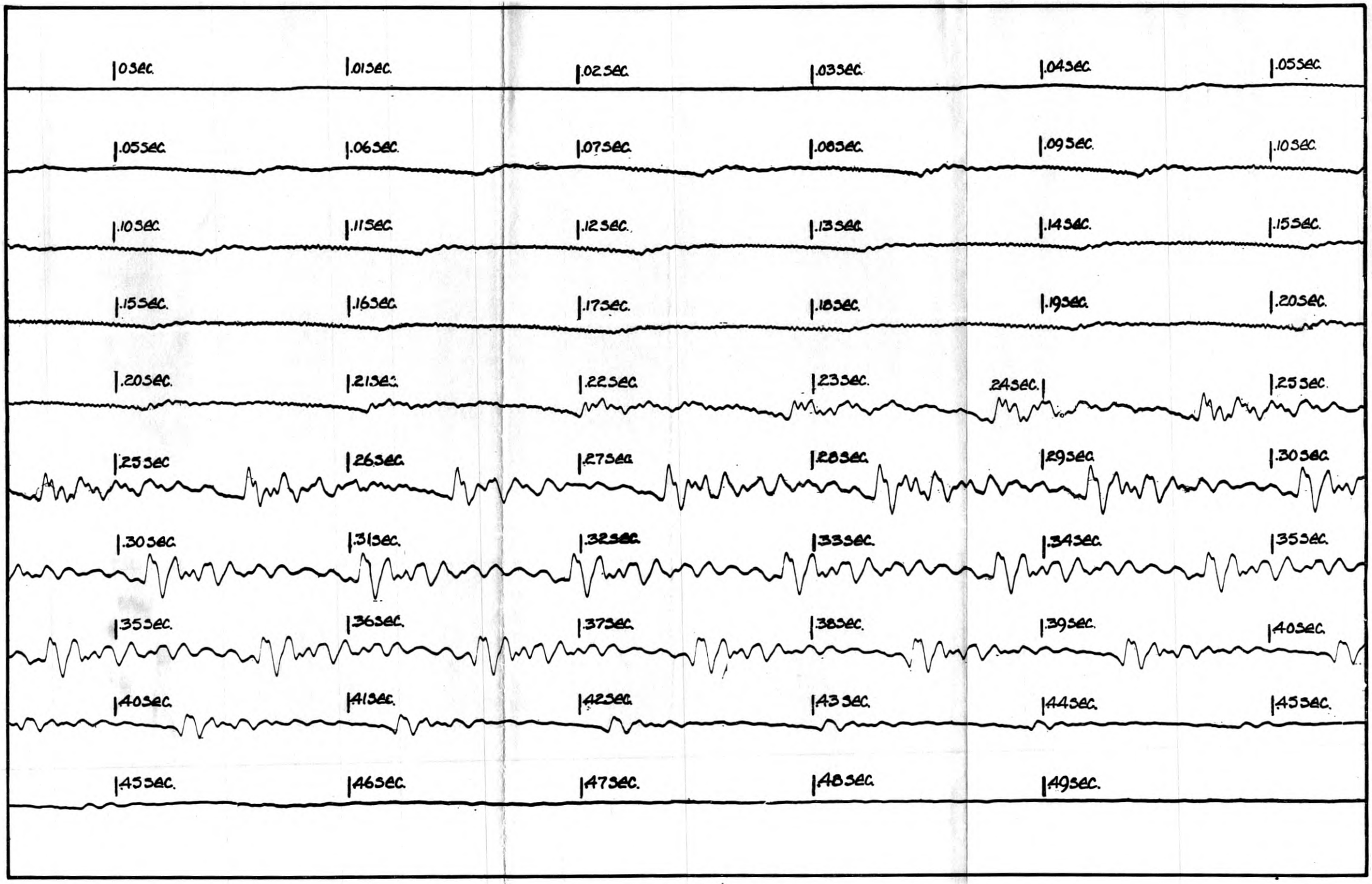


Plate No. 158—*Sa*. Spoken by M.B.

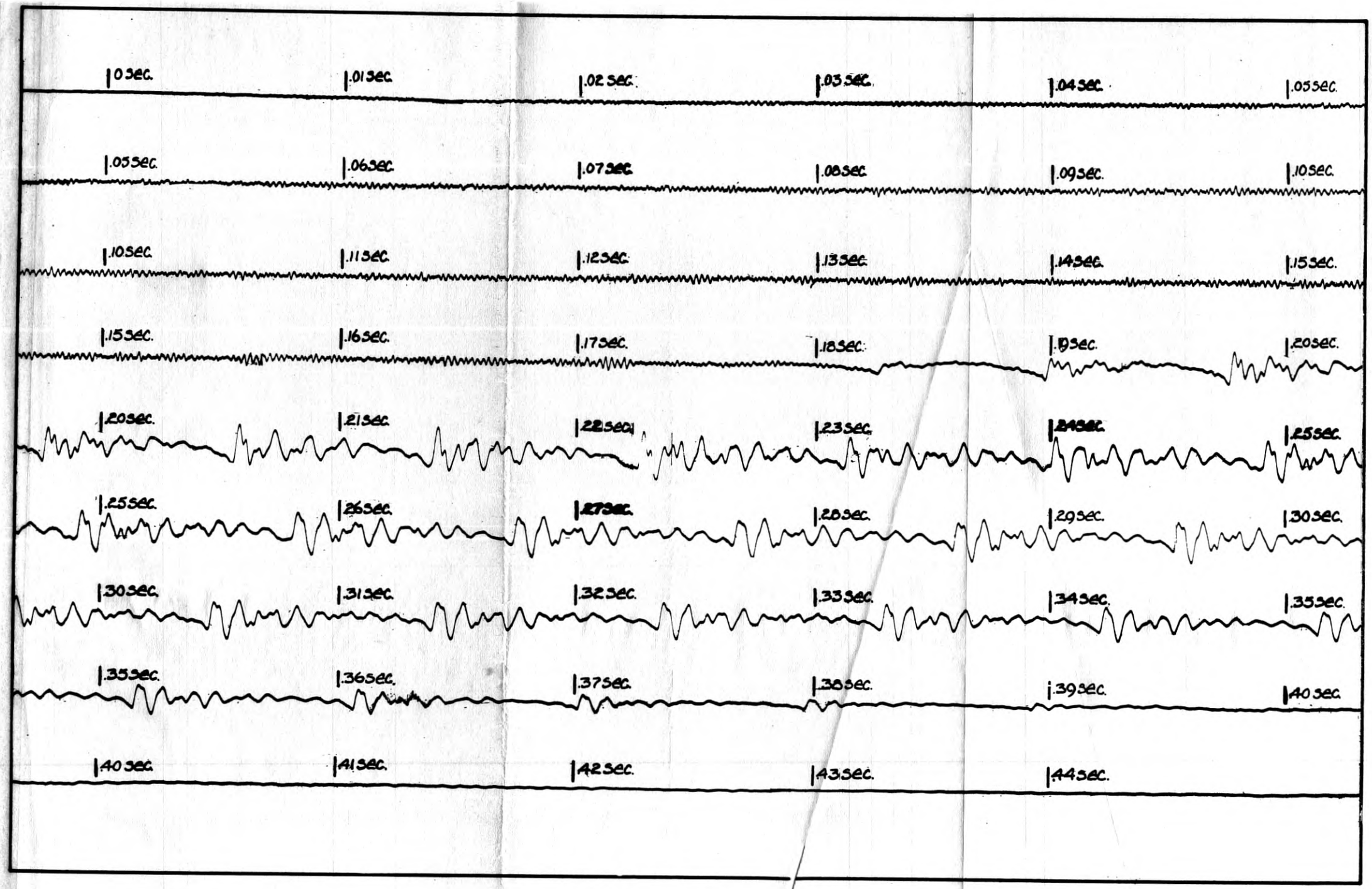


Plate No. 160—*Sa*. Spoken by M.B.

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

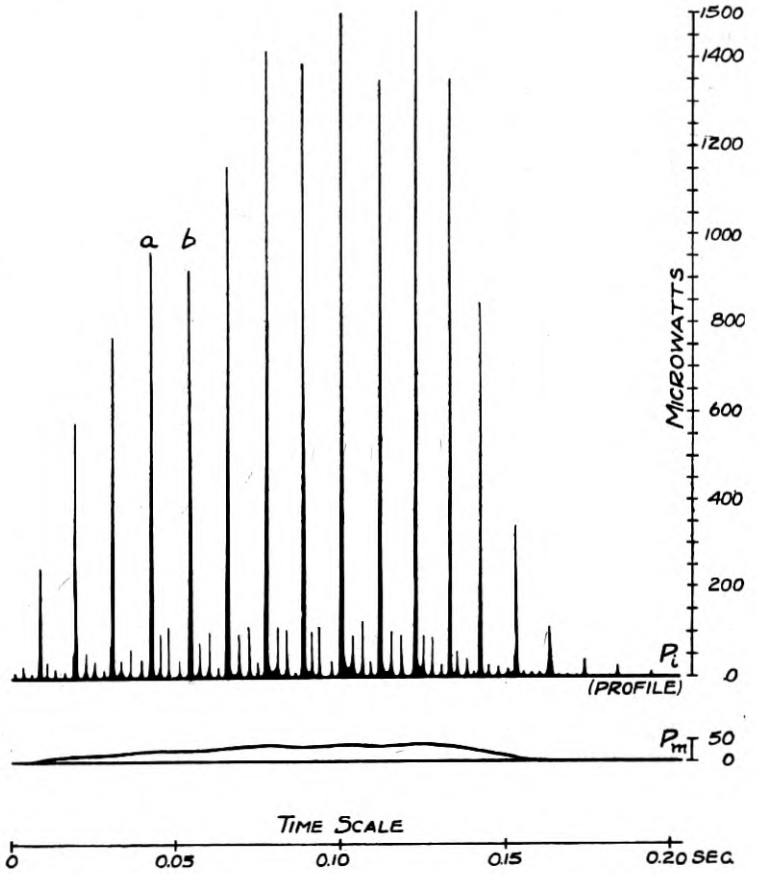


Fig. 1—Instantaneous and mean power. Enlarged copy of original oscillogram of the word "QUITE"

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 $\frac{\pi 9^2}{2} \equiv 127$. 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 over-

loaded 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 cm² at Trans.	Speaker's Power	Per cm² 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 cm² at Trans.	Speaker's Power	Per cm² 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.

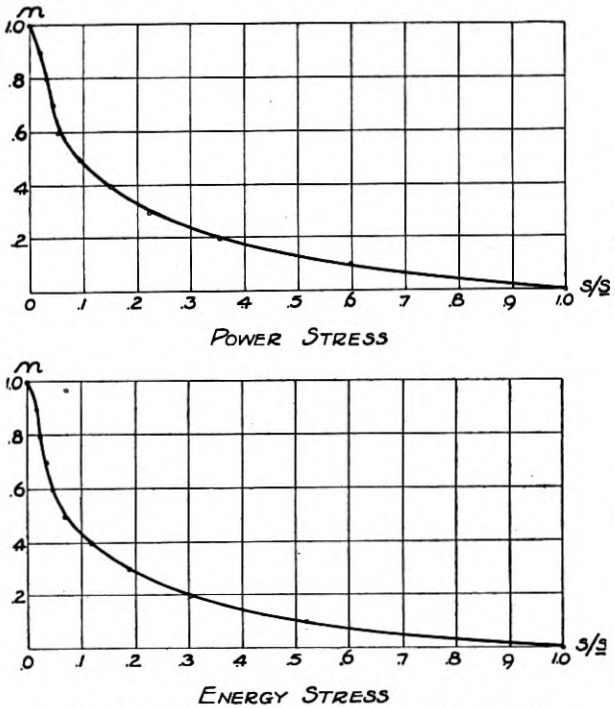


Fig. 2a—Composite stress curves of 16 voices

The analogous relation between syllabic energy and stress is found by using the total energy of each syllable instead of 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

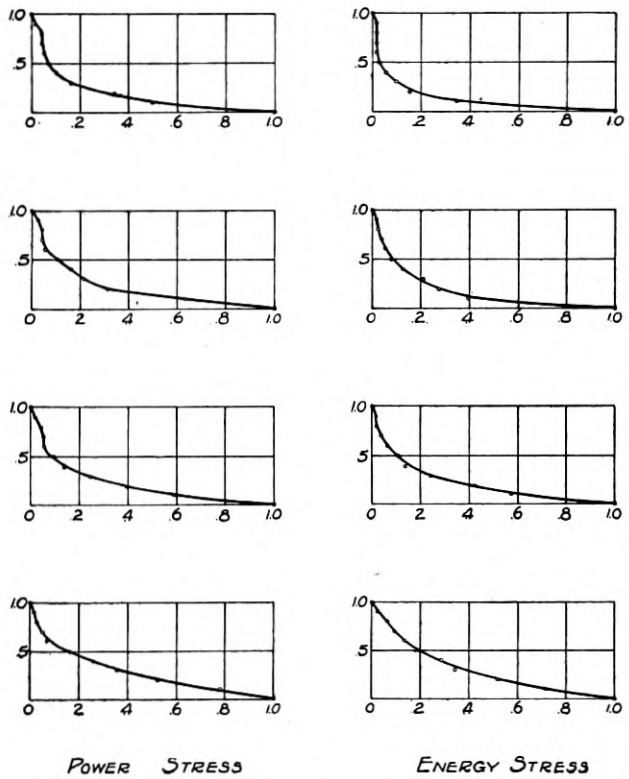


Fig. 2b—Types of stress curves

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

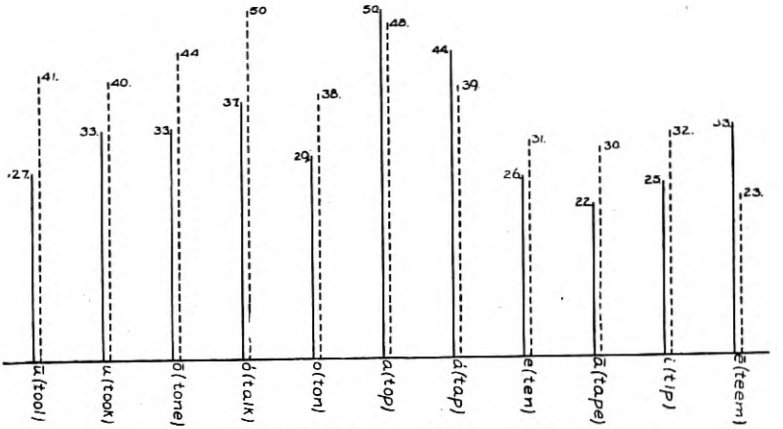


Fig. 3—Inherent relative power

——— 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 Vietor.

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

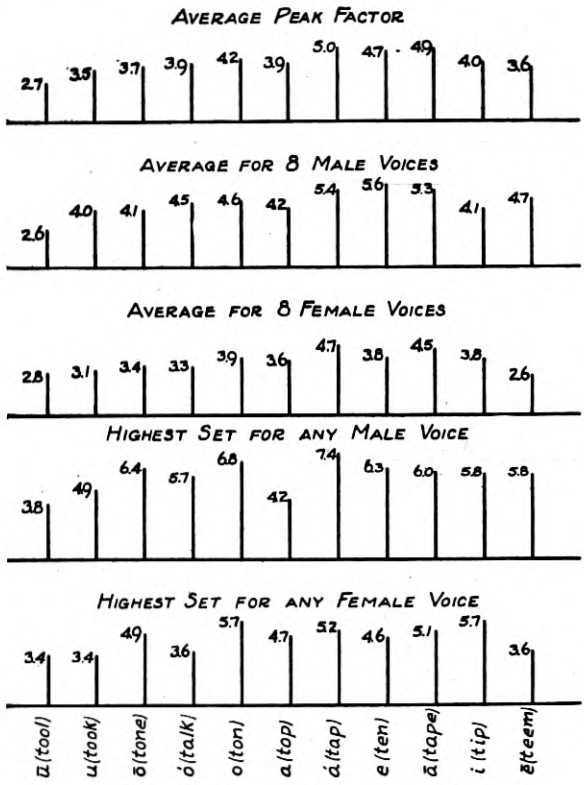


Fig. 4—Peak factors of vowels

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.

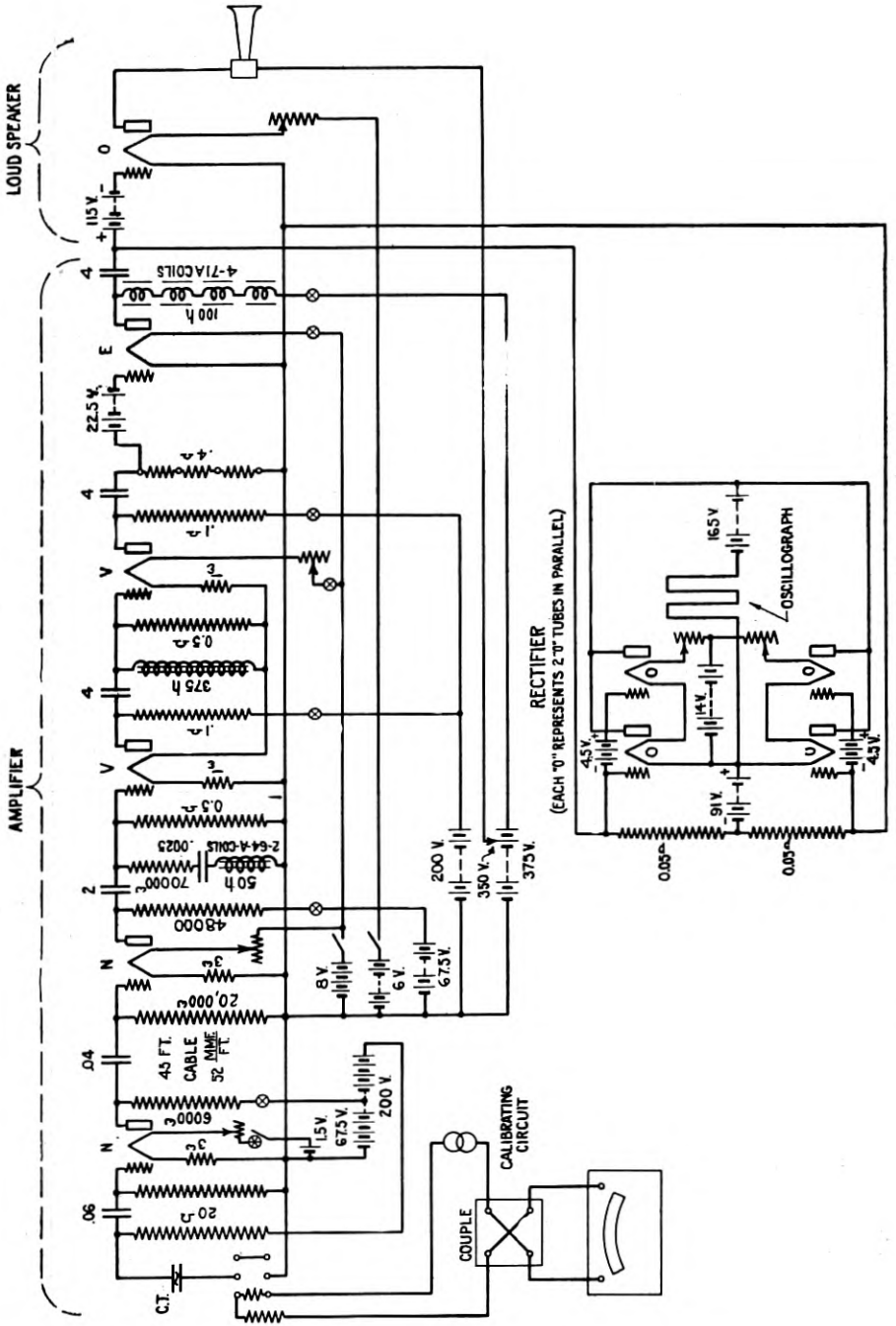


Fig. 5—Speech power recording circuit

TABLE IV

Calibration Constants

(a) Constants of Vibrators $I/D =$
  (1) Low frequency ..... 5 { milliamperes
  (2) Instantaneous power ..... 286 { per cm.
(b) Rectifier constant $E^2/I = /40$ (volts)^2/milliamp.
(c) Pressure on transmitter vs. amplifier output $p^2/E^2 = 1/2.95^2$ dynes^2/cm^4 volt^2.
(d) Power intensity at transmitter vs. pressure $P/p^2 = 1/415$ cm^2 microwatts/dynes^2.

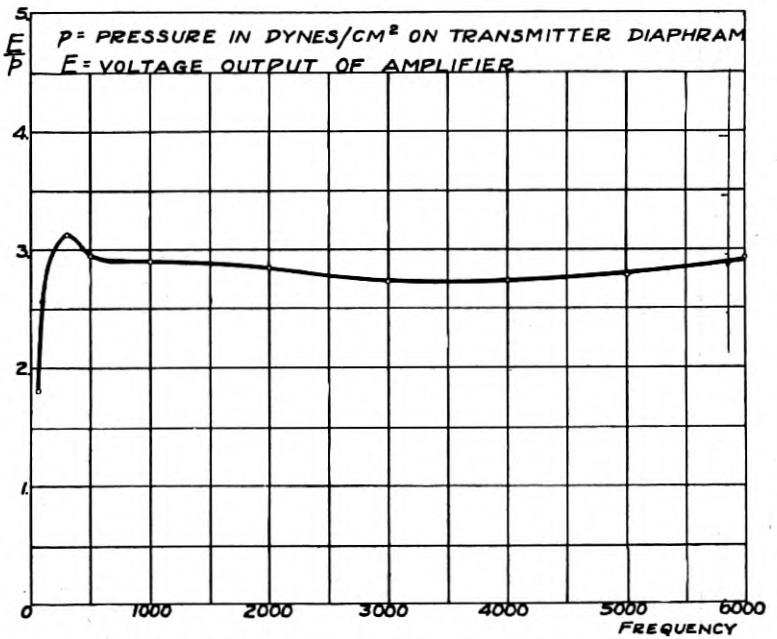


Fig. 6—Calibration of condenser transmitter with amplifier

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 ($\bar{P}_i$ and $\bar{P}_m$, 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 $f(v_o t - r) \equiv \sin \omega \left(t - \frac{r}{v_o}\right)$,

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 $\frac{v_o}{\omega}$, 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

$$h\nu_i = h\nu_{lim} - hf_i. \qquad (2)$$

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.

H. 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

$$\begin{aligned} -e^2/r &= -Rh \quad \text{for the first (or normal) stationary state} \\ -e^2/r &= -Rh/4 \quad \text{for the second stationary state} \\ -e^2/r &= -Rh/9 \quad \text{for the third stationary state; and so forth.} \end{aligned} \tag{3}$$

[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

$$\begin{aligned} r &= e^2/Rh \text{ for the normal state} \\ r &= 4e^2/Rh \text{ for the second stationary state} \\ r &= 9e^2/Rh \text{ for the third stationary state; and so on.} \end{aligned} \tag{4}$$

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

$$r = e^2/Rh,\ 4e^2/Rh,\ 9e^2/Rh, \ldots\ldots\ldots, \tag{5}$$

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

$$-e^2/2r = -Rh/n^2 \ (n = 1, 2, 3, 4, \ldots\ldots) \tag{6}$$

Each of these equations defines a value of r, as follows:

$$r = n^2e^2/2Rh \ (n = 1, 2, 3, 4, \ldots\ldots) \tag{7}$$

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

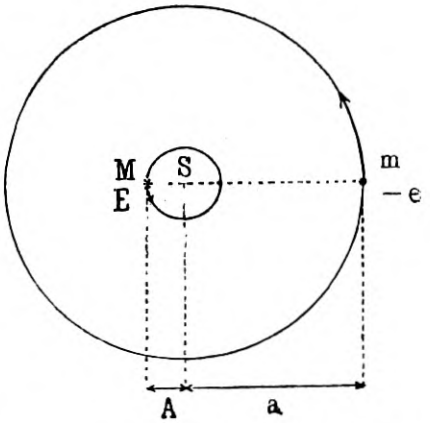


Fig. 1—Diagram to illustrate how the electron and the nucleus revolve around their common centre of mass in synchronous orbits

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 there-

fore 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

$$-e^2\mu/2a = -Rh/n^2. \tag{10}$$

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:

$$p = mva + MVA = mva/\mu. \tag{11}$$

I leave it again to the reader to use the foregoing statements to arrive at the expression

$$p = e\sqrt{ma} \tag{12}$$

and by combining (12) and (10), at the expression

$$p_n = ne^2\sqrt{m\mu/2Rh} \tag{13}$$

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

$$p_1 = e^2\sqrt{m\mu/2Rh}. \tag{14}$$

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:

$$p_1 = h/2\pi;\ p_n = nh/2\pi. \tag{15}$$

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:

$$R = 2\pi^2\mu me^4/h^3. \tag{16}$$

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* $h/2\pi$. 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 (14) 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

$$\nu = \frac{R}{n'^2 n''^2}(n'-n'')(n'+n''). \tag{20}$$

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:

$$\omega' = v/2\pi r = \sqrt{8R^3h^3}/2\pi n'^3 e^2\sqrt{m\mu} \tag{22}$$

Comparing this expression for ω' with the expression for $Lim\ \nu$ in (21), we see that they are identical, if

$$R = 2\pi^2 m\mu e^4/h^3$$

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

$$R = \frac{2\pi^2 \mu m e^4}{h^3} \tag{25}$$

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

$$R_{He}/R_H = \mu_{He}/\mu_H = (1+m/M_H)\ (1+m/M_{He}) \tag{26}$$

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 (26) is equal to 1.000403, and the ratio of the frequencies of corresponding lines in the spectra of ionized helium and of hydrogen is

$$4\ R_{He}/R_H \text{ calculated} = 4.001612 \tag{27}$$

The values of R_{He} and R_H deduced from frequency measurements yield the ratio

$$4\ R_{He}/R_H \text{ observed} = 4.0016212 \tag{28}$$

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

$$M_H/m = 1847. \tag{29}$$

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.[9]

J1. 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

$$(\zeta+a\epsilon)^2/a^2+y^2/b^2=1$$

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

$$\zeta=r\cos\phi \qquad y=r\sin\phi$$

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

$$\left(\frac{dr}{d\phi}\right)^2=\frac{r^4\epsilon^2\sin^2\phi}{a^2(1-\epsilon^2)^2}=-\frac{r^4}{a^2(1-\epsilon^2)}+\frac{2r^3}{a(1-\epsilon^2)}-r^2. \tag{31}$$

[9] 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-

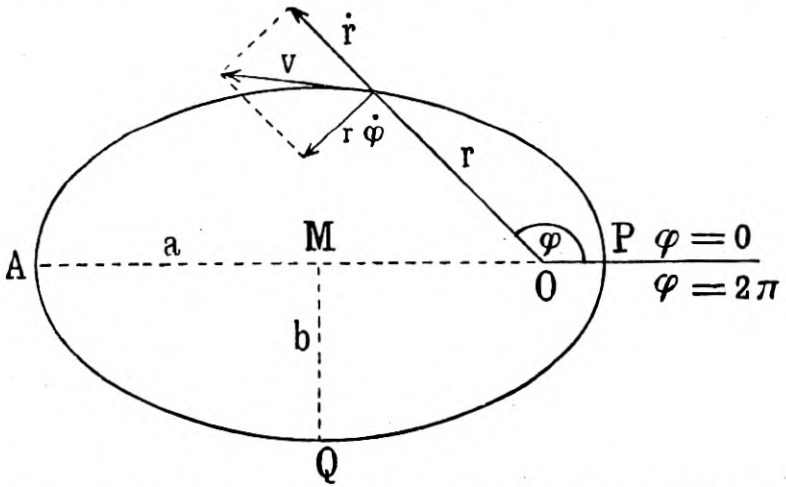


Fig. 2—Diagram to illustrate the notation used in describing elliptical orbits

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

$$\left(\frac{dr}{dt}\right)^2 = -p^2/m^2r^2+2eE/mr-C, \tag{35}$$

the last symbol standing for a constant of integration. Finally

$$(dr/d\phi)^2=(dr/dt)^2/(d\phi/dt)^2=(dr/dt)^2(m^2r^4/p^2)$$
$$=-Cmr^4/p^2+2eEmr^3/p^2-r^2. \tag{36}$$

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

$$p^2=eEma\ (1-\epsilon^2). \tag{37}$$

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

$$C=p^2/ma^2(1-\epsilon^2)=eE/a \tag{38}$$

to determine the constant of integration in (35). If now the reader will take the expression for the energy of the system

$$W=\tfrac{1}{2}mv^2-e^2/r=\tfrac{1}{2}m((dr/dt)^2+r^2(d\phi/dt))^2-e^2/r \tag{39}$$

and substitute for $(d\phi/dt)$ according to (33) and for (dr/dt) according to (35) and (38), he should arrive at

$$W=-e^2/2a. \tag{40}$$

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

$$t=\int(-p^2/m^2r^2+2eE/mr-eE/a)^{-\frac{1}{2}}dr \tag{41}$$

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

$$T_r : T_\phi = \frac{2\pi+\Delta\phi}{2\pi} = \frac{2\pi+2\pi\omega T_r}{2\pi} \tag{44}$$

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

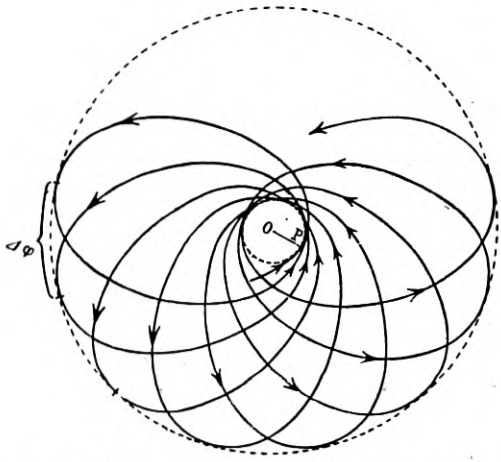


Fig. 3—Rosette orbit, resulting from a precession superposed upon an elliptical orbit

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

$$\omega_L = eH/4\pi mc \tag{48}$$

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.

J5. 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

$$I=\int 2Kdt=2\pi\sqrt{ameE} \qquad 51$$

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

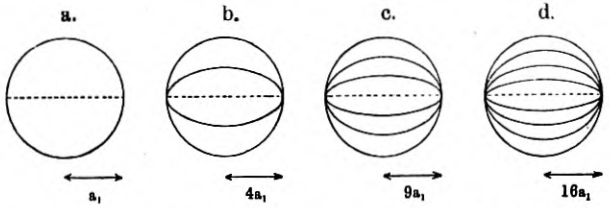


Fig. 4a—Diagram to show the proportional dimensions of ellipses with identical total quantum-number $n=I/h$ and different azimuthal quantum-numbers $k=1, 2 \ldots n$ From left to right we have the cases $n=1, 2, 3, 4$, on scales varying as indicated by the subjoined arrows.

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

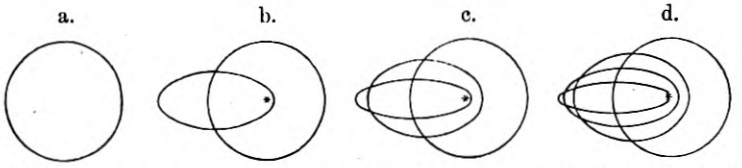


Fig. 4b—The same ellipses as appear in Fig. 4a, drawn confocally as they should appear, instead of concentrically

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 $-Rh$, $-Rh/4$, $-Rh/9 \ldots$. 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 $2h$?

$$\int p_\phi \, d\phi = 2h. \tag{61}$$

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

$$\alpha = 2\pi e^2/hc = 7.29\ 10^{-3}. \tag{63}$$

(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

$$\int p_\phi \, d\phi = h. \tag{64}$$

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 $\int p_\phi d\phi = h$, then also $\int p_r dr = h$.

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:

$$\int p_\phi d\phi = kh, \int p_r dr = (n-k)h. \tag{67}$$

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 . . . 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 . . . 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

$$\int p_\phi d\phi = nh \tag{69}$$

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 variaton 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!) 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 J4, 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 mag-

netic 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 [15].

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

$$\Delta U = MH \cos \theta \tag{77}$$

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

$$seh/4\pi mc = M \cos \theta \tag{78}$$

into which we then insert the expression for M in terms of p:

$$sh/2\pi c = p \cos \theta \tag{79}$$

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

$$s = k \cos \theta \tag{80}$$

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 $Q=\int Idt$. 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

$$\left(L_{11}\frac{d}{dt}+R_{11}+\frac{1}{C_{11}}\int dt\right)I_1+\left(L_{12}\frac{d}{dt}+R_{12}+\frac{1}{C_{12}}\int dt\right)I_2+$$
$$\ldots.+\left(L_{1n}\frac{d}{dt}+R_{1n}+\frac{1}{C_{1n}}\int dt\right)I_n=E_1$$

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

$$\left(L_{11}\frac{d}{dt}+R_{11}+\frac{1}{C_{11}}\int dt\right)I_1+\ldots+\left(L_{1n}\frac{d}{dt}+R_{1n}+\frac{1}{C_{1n}}\int dt\right)I_n=E_1$$
$$\text{- -} \qquad (1)$$
$$\left(L_{n1}\frac{d}{dt}+R_{n1}+\frac{1}{C_{n1}}\int dt\right)I_1+\ldots+\left(L_{nn}\frac{d}{dt}+R_{nn}+\frac{1}{C_{nn}}\int dt\right)I_n=E_n$$

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

$$\begin{array}{lll} L_{11}=L_1 & L_{22}=L_2 & L_{12}=L_{21}=+M \\ C_{11}=C_1 & C_{22}=C_2 & C_{12}=C_{21}=o \\ R_{11}=R_1 & R_{22}=R_2 & R_{12}=R_{21}=o \end{array}$$

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.

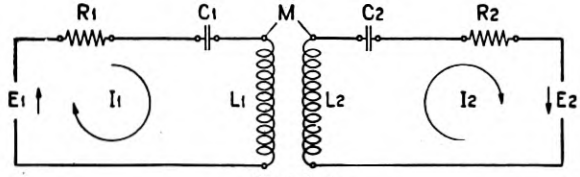
Now write down Kirchhoff's Law, or the circuital equation for the network of sketch 2. They are

$$\left\{(L_1+L_3)\frac{d}{dt}+(R_1+R_3)+\left(\frac{1}{C_1}+\frac{1}{C_3}\right)\int dt\right\} I_1$$
$$-\left(L_3\frac{d}{dt}+R_3+\frac{1}{C_3}\int dt\right) I_2=E_1,$$
$$-\left(L_3\frac{d}{dt}+R_3+\frac{1}{C_3}\int dt\right) I_1$$
$$+\left\{(L_2+L_3)\frac{d}{dt}+(R_2+R_3)+\left(\frac{1}{C_2}+\frac{1}{C_3}\right)\int dt\right\} I_2=E_2.$$

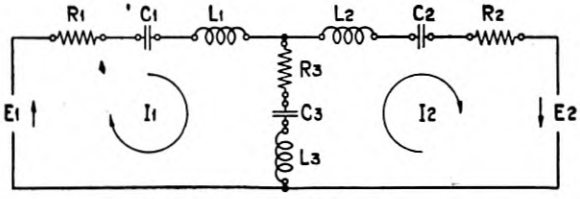
Comparison with equations (1) shows that

$$L_{11}=L_1+L_3 \qquad L_{22}=L_2+L_3 \qquad L_{12}=L_{21}=-L_3$$
$$R_{11}=R_1+R_3 \qquad R_{22}=R_2+R_3 \qquad R_{12}=R_{21}=-R_3$$
$$\frac{1}{C_{11}}=\frac{1}{C_1}+\frac{1}{C_3} \qquad \frac{1}{C_{22}}=\frac{1}{C_2}+\frac{1}{C_3} \qquad \frac{1}{C_{12}}=\frac{1}{C_{21}}=-\frac{1}{C_3}.$$

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

$$\frac{d}{dt}\sum\sum\frac{1}{2}L_{jk}I_jI_k+\frac{d}{dt}\sum\sum\frac{1}{2}\frac{1}{C_{jk}}Q_jQ_k+\sum\sum R_{jk}I_jI_k=\sum E_jI_j. \tag{2}$$

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, $\sum\sum R_{jk}I_jI_k$, 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

$$\begin{aligned}&\left(L_{11}\frac{d^2}{dt^2}+R_{11}\frac{d}{dt}+S_{11}\right)Q_1+\ldots+\left(L_{1n}\frac{d^2}{dt^2}+R_{1n}\frac{d}{dt}+S_{1n}\right)Q_n=E_1,\\&\text{-----------------------------------}\\&\left(L_{n1}\frac{d^2}{dt^2}+R_{n1}\frac{d}{dt}+S_{n1}\right)Q_1+\ldots+\left(L_{nn}\frac{d^2}{dt^2}+R_{nn}\frac{d}{dt}+S_{nn}\right)Q_n=E_n.\end{aligned}\tag{3}$$

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

$$\begin{aligned}&(\lambda L_{11}+R_{11}+1/\lambda C_{11})J_1+\ldots+(\lambda L_{1n}+R_{1n}+1/\lambda C_{1n})J_n=F_1,\\&(\lambda L_{21}+R_{21}+1/\lambda C_{21})J_1+\ldots+(\lambda L_{2n}+R_{2n}+1/\lambda C_{2n})J_n=o,\\&\text{-----------------------------------}\\&(\lambda L_{n1}+R_{n1}+1/\lambda C_{n1})J_1+\ldots+(\lambda L_{nn}+R_{nn}+1/\lambda C_{nn})J_n=o.\end{aligned}\tag{4}$$

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

$$\begin{aligned} &z_{11}J_1+z_{12}J_2+\ldots+z_{1n}J_n=F_1,\\ &z_{21}J_1+z_{22}J_2+\ldots+z_{2n}J_n=0,\\ &\text{- - - - - - - - - - - - - - - - - -}\\ &z_{n1}J_1+z_{n2}J_2+\ldots+z_{nn}J_n=0. \end{aligned} \tag{5}$$

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,

$$\begin{vmatrix} z_{11} & z_{12} & z_{13} & \ldots\ldots & z_{1n}\\ z_{21} & z_{22} & z_{23} & \ldots\ldots & z_{2n}\\ z_{31} & z_{32} & \ldots\ldots & & z_{3n}\\ \text{- - - - - - - - - - - - - -} & & & & \\ z_{n1} & z_{n2} & \ldots\ldots & & z_{nn} \end{vmatrix} \tag{7}$$

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 $\lambda, i\omega$ and $-i\omega$ respectively. That is,

$$J_j'=\frac{1}{2}\frac{F'}{Z_{j1}(i\omega)}\text{ and }J_j''=\tfrac{1}{2}\frac{F''}{Z_{j1}(-i\omega)}$$

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,

$$\begin{aligned} &\left(L_{11}\frac{d}{dt}+R_{11}+\frac{1}{C_{11}}\int dt\right)I_1'+\ldots+\left(L_{1n}\frac{d}{dt}+R_{1n}+\frac{1}{C_{1n}}\int dt\right)I_n'=o, \\ &\text{-------------------------------} \qquad (14) \\ &\left(L_{n1}\frac{d}{dt}+R_{n1}+\frac{1}{C_{n1}}\int dt\right)I_1'+\ldots+\left(L_{nn}\frac{d}{dt}+R_{nn}+\frac{1}{C_{nn}}\int dt\right)I_n'=o. \end{aligned}$$

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 (14), assume that a solution exists of the form

$$I_j' = J_j' e^{\lambda t} \qquad (j=1,2 \ldots n)$$

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

$$\begin{aligned} &Z_{11}(\lambda)J_j' + \ldots . + Z_{1n}(\lambda)J_n' = o, \\ &\text{- -} \\ &Z_{n1}(\lambda)J_1' + \ldots . + Z_{nn}(\lambda)J_n' = o. \end{aligned} \tag{15}$$

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 $\lambda_1, \lambda_2 \ldots \lambda_m$. 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

$$I_1' = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} + \ldots . + C_m e^{\lambda_m t}, \tag{17}$$

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

$$E_oA(t)+E_1A(t-\tau_1)+\ldots+E_nA(t-\tau_n)=\sum_{j=o}^{n}E_jA(t-\tau_j). \quad (18)$$

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 $E(o)+\Delta_1 E$ for $\Delta t\leq t<2\Delta t$.
4. It has the value $E(o)+\Delta_1 E+\Delta_2 E$ for $2\Delta t\leq t<3\Delta t$.

In other words it has the increment $\Delta_j E$ at time $t=j\Delta t$.

Evidently then the resultant current $I(t)$ is

$$E_oA(t)+\Delta_1 EA(t-\Delta t)+\ldots+\Delta_n E.A(t-n\Delta t).$$

Now evidently if the interval Δt is made shorter and shorter, then in the limit $\Delta t\to dt$ and $j\Delta t=\tau$ and

$$\Delta_j E=\frac{d}{d\tau}E(\tau)d\tau.$$

Passing to the limit in the usual manner this summation becomes a definite integral and we get

$$I(t)=E(o)A(t)+\int_0^t A(t-\tau)\frac{d}{d\tau}E(\tau)d\tau. \quad (19)$$

Finally by obvious transformations of the expression we arrive at the fundamental formula of circuit theory

$$I(t)=\frac{d}{dt}\int_0^t A(t-\tau)E(\tau)d\tau, \quad (20)$$

$$=\frac{d}{dt}\int_0^t E(t-\tau)A(\tau)d\tau. \quad (20\text{-a})$$

For completeness we write down the following equivalents of (20) and (20-a)

$$I(t) = A(o)E(t) + \int_0^t A'(t-\tau)E(\tau)d\tau, \tag{20-b}$$

$$= A(o)E(t) + \int_0^t A'(\tau)E(t-\tau)d\tau, \tag{20-c}$$

$$= E(o)A(t) + \int_0^t E'(t-\tau)A(\tau)d\tau, \tag{20-d}$$

$$= E(o)A(t) + \int_0^t E'(\tau)A(t-\tau)d\tau. \tag{20-e}$$

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

$$\begin{aligned} I(t) &= A(o)E \sin(\omega t+\theta) \\ &\quad + E \sin(\omega t+\theta)\int_0^t \cos \omega\tau A'(\tau)d\tau \\ &\quad - E \cos(\omega t+\theta)\int_0^t \sin \omega\tau A'(\tau)d\tau \end{aligned} \tag{21}$$

where $A'(t)=\frac{d}{dt}A(t)$.

Now this can be resolved into two parts

$$\begin{aligned} &E \sin(\omega t+\theta)\left\{A(o)+\int_0^\infty \cos \omega\tau A'(\tau)d\tau\right\} \\ &-E \cos(\omega t+\theta)\left\{\int_0^\infty \sin \omega\tau A'(\tau)d\tau\right\} \end{aligned} \tag{22}$$

which is the *final steady state*, and

$$\begin{aligned} &-E \sin(\omega t+\theta)\int_t^\infty \cos \omega\tau A'(\tau)d\tau \\ &+E \cos(\omega t+\theta)\int_t^\infty \sin \omega\tau A'(\tau)d\tau \end{aligned} \tag{23}$$

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

$$\alpha(\omega)=A(o)+\int_0^\infty \cos \omega\tau\ A'(\tau)d\tau, \tag{24}$$

$$\beta(\omega)=-\int_0^\infty \sin \omega\tau\ A'(\tau)d\tau. \tag{25}$$

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

$$I(t)=\frac{e^{pt}}{Z(p)}+y(t). \tag{26}$$

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

$$\frac{d}{dt}\left\{e^{pt}\int_0^\infty A(\tau)e^{-t\tau}d\tau-e^{pt}\int_t^\infty A(\tau)e^{-p\tau}d\tau\right\}.$$

Carrying out the indicated differentiation this becomes

$$I(t)=pe^{pt}\int_0^\infty A(\tau)e^{-p\tau}d\tau-pe^{pt}\int_t^\infty A(\tau)e^{-p\tau}d\tau+A(t). \tag{27}$$

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,

$$A = \frac{1}{pL}\frac{1}{1+\alpha/p} = \frac{1}{R}\frac{\alpha}{p}\frac{1}{1+\alpha/p} = \frac{1}{R}\left\{\frac{\alpha}{p} - (\frac{\alpha}{p})^2 + (\frac{\alpha}{p})^3 - (\frac{\alpha}{p})^4 + \cdots\right\}$$

by the Binomial Theorem.

Now expand the explicit solution as a power series in t: it is

$$A=\frac{1}{R}\left\{\frac{\alpha t}{1!}-\frac{(\alpha t)^2}{2!}+\frac{(\alpha t)^3}{3!}-\cdots\right\}.$$

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

$$A=\frac{1}{R}\left\{1-(p/\alpha)+(p/\alpha)^2-\cdots\right\}$$

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

$$C+k_1+k_2=o,$$
$$\lambda_1k_1+\lambda_2k_2=o,$$

and

$$k_1=\lambda_2C/(\lambda_1-\lambda_2),$$
$$k_2=\lambda_1C/(\lambda_2-\lambda_1).$$

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

$$Q=C+k_1e^{\lambda_1t}+k_2e^{\lambda_2t},$$

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*:—

$$h = a_0 + a_1 t/1! + a_2 t^2/2! + \ldots + a_n t^n/n! + \ldots \tag{35}$$

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

$$h_0 + h_1 t/1! + h_2 t^2/2! + \ldots .$$

Substitute this assumed expansion in the integral, and integrate term by term. The right hand side of the integral equation becomes formally

$$h_0/p + h_1/p^2 + h_2/p^3 + \ldots .$$

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

$$a_0/p + a_1/p^2 + a_2/p^3 + \ldots .$$

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

$$h(t) = a_o + a_1 t/1! + a_2 t^2/2! + \ldots .$$

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

$$h(t) = a_o + a_1 t/1! + a_2 t^2/2! + \ldots = \sum_{o}^{\infty} a_n t^n/n! \qquad (35)$$

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

$$a_o + a_1 t/1! + a_2 t^2/2! + \ldots.$$

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

$$G(x)=G(o)+G^{(1)}(o)\frac{x}{1!}+G^{(2)}(o)\frac{x^2}{2!}+\cdots$$

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

$$G(x)=1/H(p)=a_o+a_1/p+a_2/p^2+\cdots$$

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

$$\frac{1}{pH(p)} = \int_0^\infty h(t)e^{-pt}dt$$

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

$$H'(p_j) = \left\{ \frac{d}{dp} H(p) \right\}_{p=p_j}. \tag{41}$$

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

$$h_o(t) = h_o = \frac{1}{H(o)}, \tag{47}$$

$$h_j(t) = \frac{e^{p_j t}}{p_j H'(p_j)}, \qquad j = 1,2, \ldots n.$$

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

$$C_o H(p) + \sum \frac{pH(p)}{p-p_j} C_j = 1. \tag{51}$$

If now $p=p_j+q$ and q approaches zero, this becomes in the limit

$$p_j H'(p_j) C_j = 1 \tag{52}$$

or

$$C_j = \frac{1}{p_j H'(p_j)}, \tag{53}$$

whence

$$h(t) = \frac{1}{H(o)} + \sum \frac{e^{p_j t}}{p_j H'(p_j)} \tag{54}$$

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

$$Q = \frac{1}{L}\frac{1}{p^2+2\alpha p+\omega^2}$$

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

$$Q = C - \frac{e^{-\alpha t}}{\omega' L}\,\frac{\omega' \cos \omega' t + \alpha \sin \omega' t}{\alpha^2 + \omega'^2}.$$

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) $$\int_\lambda^\infty e^{-pt} J_o\left(\sqrt{t^2-\lambda^2}\right) dt = \frac{e^{-\lambda\sqrt{p^2+1}}}{\sqrt{p^2+1}},$$

$$h = \frac{p}{\sqrt{p^2+1}} e^{-\lambda\sqrt{p^2+1}} = o \text{ for } t<\lambda$$

$$= J_o\left(\sqrt{t^2-\lambda^2}\right) \text{ for } t \geq \lambda.$$

(n) $$\int_0^\infty e^{-pt} J_n(\lambda t) dt = \frac{1}{r}\left(\frac{r-p}{\lambda}\right)^n, \; r^2 = p^2+\lambda^2,$$

$$h = \frac{p}{r}\left(\frac{r-p}{\lambda}\right)^n = J_n(\lambda t).$$

(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), $I_o(x)$ denotes the Bessel function $J_o(ix)$ 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

$$F(p)=\int_0^\infty f(t)e^{-pt}dt$$

$$F_1(p)=\int_0^\infty f_1(t)e^{-pt}dt$$

$$F_2(p)=\int_0^\infty f_2(t)e^{-pt}dt$$

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

$$g(t)=(1+\lambda\int_a^t dt)e^{-\lambda t}h(t).$$

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

$$\begin{aligned} RI &= -\frac{\partial}{\partial x} V \\ C\frac{d}{dt} V &= -\frac{\partial}{\partial x} I \end{aligned} \tag{57}$$

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

$$\begin{aligned} RI &= -\frac{\partial}{\partial x} V \\ pCV &= -\frac{\partial}{\partial x} I. \end{aligned} \tag{58}$$

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

$$V = V_1 e^{-\gamma x} + V_2 e^{\gamma x} \tag{59}$$

$$I = \sqrt{\frac{pC}{R}} [V_1 e^{-\gamma x} - V_2 e^{\gamma x}] \tag{60}$$

where

$$\gamma = \sqrt{pRC}. \tag{61}$$

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,

$$V = Ee^{-x\sqrt{pCR}} = Ee^{-\sqrt{\alpha p}} \tag{62}$$

$$I = \sqrt{\frac{pC}{R}} Ee^{-x\sqrt{pCR}} = \sqrt{\frac{pC}{R}} Ee^{-\sqrt{\alpha p}} \tag{63}$$

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

$$V = e^{-\sqrt{\alpha p}} \tag{64}$$

$$I = \sqrt{\frac{pC}{R}} e^{-\sqrt{\alpha p}}. \tag{65}$$

Now suppose that $x=o$ so that $\alpha=o$, in other words consider a point at the cable terminals. Then

$$V = 1$$

$$I = \sqrt{\frac{pC}{R}}. \tag{66}$$

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

$$I = \sqrt{\frac{pC}{R}}. \tag{66}$$

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,

$$\sqrt{\frac{C}{R}}\frac{\sqrt{p+\lambda}}{p}=\int_0^\infty I(t)e^{-pt}dt. \tag{72}$$

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

$$\int^t\frac{e^{-\lambda t}}{\sqrt{t}}dt=2\int_0^t e^{-\lambda t}d\sqrt{t}=2\sqrt{t}\,e^{-\lambda t}+2\lambda\int_0^t e^{-\lambda t}\sqrt{t}\,dt.$$

Proceeding in this way by repeated partial integration we get for the integral term of (77)

$$2\sqrt{t}\,e^{-\lambda t}\left\{1+\frac{2\lambda t}{1.3}+\frac{(2\lambda t)^2}{1.3.5}+\ldots\right\}. \tag{78}$$

The straightforward Heaviside solution is obtained by expanding the operational equation as follows:

$$I=\sqrt{\frac{C}{R}}\sqrt{p+\lambda}$$

$$=\sqrt{\frac{C}{R}}\left(1+\frac{\lambda}{p}\right)^{1/2}\sqrt{p}$$

$$=\sqrt{\frac{C}{R}}\left[1+\frac{1}{2}\frac{\lambda}{p}-\frac{1}{2.4}\left(\frac{\lambda}{p}\right)^2+\frac{1.3}{2.4.6}\left(\frac{\lambda}{p}\right)^3-\cdots\right]\sqrt{p}.$$

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

$$I=\sqrt{\frac{C}{\pi Rt}}\left\{1+\frac{(2\lambda t)}{2}-\frac{(2\lambda t)^2}{2.3.4}+\frac{1.3(2\lambda t)^3}{2.3.4.5.6}-\cdots\right\}. \tag{79}$$

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

$$\begin{aligned}\int_t^\infty \frac{e^{-\lambda t}}{\sqrt{t}}dt = \frac{e^{-\lambda t}}{\lambda\sqrt{t}}\Big[1 - \frac{1}{2\lambda t} + \frac{1.3}{(2\lambda t)^2} - \frac{1.3.5}{(2\lambda t)^3} \\ + \cdots + (-1)^n \frac{1.3.5 \cdots (2n-1)}{(2\lambda t)^n}\Big] \\ - \frac{(-1)^n}{\lambda}\frac{1.3.5 \cdots (2n+1)}{2(2\lambda)^n}\int_t^\infty \frac{e^{-\lambda t}}{t^{n+1}\sqrt{t}}dt.\end{aligned} \tag{82}$$

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*

$$\frac{(-1)^n}{\lambda\sqrt{t}}\frac{1.3.5 \cdots (2n-1)}{(2\lambda t)^{n-1}}e^{-\lambda t}. \tag{83}$$

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

$$I \infty \sqrt{\frac{\lambda C}{R}} + \sqrt{\frac{C}{\pi R t}}e^{-\lambda t}\left\{\frac{1}{2\lambda t} - \frac{1.3}{(2\lambda t)^2} + \frac{1.3.5}{(2\lambda t)^3} - \cdots\right\}. \tag{84}$$

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

$$\begin{aligned} V &= 1-\sqrt{\frac{a}{p}}+\frac{a}{p}-\frac{a}{p}\sqrt{\frac{a}{p}}+\left(\frac{a}{p}\right)^2-\cdots \\ &= 1+\frac{a}{p}+\left(\frac{a}{p}\right)^2+\cdots. \\ &\quad -\left(1+\frac{a}{p}+\left(\frac{a}{p}\right)^2+\cdots\right)\sqrt{\frac{a}{p}}, \\ &= 1+\frac{at}{1!}+\frac{(at)^2}{2!}+\cdots. \\ &\quad -\left(\frac{2at}{1}+\frac{(2at)^2}{1.3.}+\frac{(2at)^3}{1.3.5.}+\cdots\right)\frac{1}{\sqrt{\pi at}} \end{aligned} \tag{86}$$

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

$$V(t)=\sqrt{\frac{a}{\pi}}e^{at}\int^\infty\frac{e^{-a\tau}}{\sqrt{\tau}}\,d\tau. \tag{89}$$

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

$$\int_t^\infty \frac{e^{-a\tau}}{\sqrt{\tau}}\,d\tau = -\frac{1}{a}\int_t^\infty \frac{1}{\sqrt{\tau}} d\,e^{-a\tau}$$

$$= \frac{e^{-at}}{a\sqrt{t}} - \frac{1}{2a}\int_t^\infty \frac{e^{-a\tau}}{\tau\sqrt{\tau}}\,d\tau$$

$$= \frac{e^{-at}}{a\sqrt{\tau}} + \frac{1}{2a^2}\int_t^\infty \frac{1}{\tau\sqrt{\tau}} d\,e^{-a\tau}$$

$$= \frac{e^{-at}}{a\sqrt{t}} - \frac{e^{-at}}{2a^2t\sqrt{t}} + \frac{1.3}{2^2a^2}\int_t^\infty \frac{e^{-a\tau}}{\tau^2\sqrt{\tau}}\,d\tau$$

and finally

$$\frac{e^{-at}}{a\sqrt{t}}\left\{1 - \frac{1}{2at} + \frac{1.3}{(2at)^2} - \frac{1.3.5}{(2at)^3} + \cdots\right\}. \tag{90}$$

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

$$V(t) \sim \frac{1}{\sqrt{\pi at}}\left\{1 - \frac{1}{2at} + \frac{1.3}{(2at)^2} - \frac{1.3.5}{(2at)^3} + \cdots\right\}. \tag{91}$$

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

$$V = \left\{1 - \sqrt{\frac{p}{a}} + \frac{p}{a} - \frac{p}{a}\sqrt{\frac{p}{a}} + \left(\frac{p}{a}\right)^2 - \cdots\right\}\sqrt{\frac{p}{a}}$$

$$= \left(1 + \frac{p}{a} + \left(\frac{p}{a}\right)^2 + \cdots\right)\sqrt{\frac{p}{a}} \tag{93}$$

$$- \left(\frac{p}{a} + \left(\frac{p}{a}\right)^2 + \left(\frac{p}{a}\right)^3 + \cdots\right).$$

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

$$V=\left(1+\frac{1}{a}\frac{d}{dt}+\frac{1}{a^2}\frac{d^2}{dt^2}+\cdots\right)\frac{1}{\sqrt{\pi at}} \tag{94}$$

or, carrying out the indicated differentiation,

$$V=\frac{1}{\sqrt{\pi at}}\left(1-\frac{1}{2at}+\frac{1.3}{(2at)^2}-\frac{1.3.5}{(2at)^3}+\cdots\right)$$

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

$$\begin{aligned}\left(L\frac{d}{dt}+R\right)I&=-\frac{\partial}{\partial x}V\\ C\frac{d}{dt}V&=-\frac{\partial}{\partial x}I.\end{aligned} \tag{95}$$

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

$$I=\sqrt{\frac{pC}{pL+R}}V_0 \tag{97}$$

where V_0 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

$$I=\sqrt{\frac{C}{L}}e^{-\lambda t}I_0(\lambda t) \tag{101}$$

where $I_0(\lambda t)$ is the Bessel function $J_0(i\lambda t)$, 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

$$I=\sqrt{\frac{C}{L}}\left\{1-\frac{\lambda}{p}+\frac{1.3}{2!}\left(\frac{\lambda}{p}\right)^2-\frac{1.3.5}{3!}\left(\frac{\lambda}{p}\right)^3+\cdots\right\} \tag{102}$$

and replaces $1/p^n$ by t^n/n in accordance with the rule discussed in preceding sections. The explicit solution is then

$$I=\sqrt{\frac{C}{L}}\left\{1-\lambda t+\frac{1.3}{(2!)^2}(\lambda t)^2-\frac{1.3.5}{(3!)^2}(\lambda t)^3+\cdots\right\} \tag{103}$$

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

$$I=\sqrt{\frac{C}{L}}\left\{1-\frac{p}{4\lambda}+\frac{1.3}{2!}\left(\frac{p}{4\lambda}\right)^2-\frac{1.3.5}{3!}\left(\frac{p}{4\lambda}\right)^3+\cdots\right\}\sqrt{\frac{p}{2\lambda}}. \tag{105}$$

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

$$I=\sqrt{\frac{C}{L}}\,\frac{1}{\sqrt{2\pi\lambda t}}\left\{1+\frac{1}{8\lambda t}+\frac{(1.3)^2}{2!(8\lambda t)^2}+\frac{(1.3.5)^2}{3!(8\lambda t)^3}+\cdots\right\}. \tag{106}$$

This series solution is divergent: Heaviside recognizes it, however, as the asymptotic expansion of the function $e^{-\lambda t}I_o(\lambda t)$, and thus arrives at the solution

$$I=\sqrt{\frac{C}{L}}\,e^{-\lambda t}I_o(\lambda t) \tag{101}$$

which we have obtained from our tables of integrals.

Now the divergent expansion (106) is the well known asymptotic expansion of the function $e^{-\lambda t}I_o(\lambda t)$, 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

$$h = a_0 + a_1 p + a_2 p^2 + \ldots + a_n p^n + \ldots$$
$$(b_0 + b_1 p + b_2 p^2 + \ldots + b_n p^n + \ldots)\sqrt{p}, \tag{107}$$

a solution, usually divergent, is obtained by discarding the first expansion entirely, except for the leading constant terms a_0, 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.

$$h = a_0 + \left(b_0 + b_1\frac{d}{dt} + b_2\frac{d^2}{dt^2} + \ldots\right)\frac{1}{\sqrt{\pi t}} \tag{108}$$

$$= a_0 + \frac{1}{\sqrt{\pi t}}\left(b_0 - b_1\frac{1}{2t} + b_2\frac{1.3}{(2t)^2} - b_3\frac{1.3.5}{(2t)^3} + \ldots\right). \tag{109}$$

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 re-

markably 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 (III) 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

$$h(t)\sim\frac{1}{\sqrt{\pi t}}\left\{b_0-b_1\frac{1}{2t}+b_2\frac{1.3}{(2t)^2}-b_3\frac{1.3.5}{(2t)^3}+\ldots\right\} \tag{119}$$

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*

$$F(p)=b_0+b_1p+b_2p^2+\ldots+b_np^n+\ldots.$$

so that, formally,

$$h=(b_0+b_1p+b_2p^2+\ldots+b_np^n+\ldots)\sqrt{p}$$

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*

$$h(t)\sim\left(b_0+b_1\frac{d}{dt}+b_2\frac{d^2}{dt^2}+\ldots.\right)\frac{1}{\sqrt{\pi t}}$$

$$\sim\frac{1}{\sqrt{\pi t}}\left(b_0-b_1\frac{1}{2t}+b_2\frac{1.3}{(2t)^2}-b_3\frac{1.3.5}{(2t)^3}+\ldots\right)$$

provided the function $k=k(t)$, defined by the operational equation $k=F(p)$, and the infinite integrals

$$\int_0^\infty t^n k(t)dt \qquad (n=1,2,\ldots.)$$

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

$$I=\sqrt{\frac{C}{L}}e^{-\lambda t}I_o(\lambda t).$$

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

$$\frac{1}{\pi\sqrt{t}}\left\{\int_0^\infty\frac{e^{-2\lambda t}}{\sqrt{t}}dt+\frac{1}{2t}\int_0^\infty\frac{\sqrt{t}}{1!}e^{-2\lambda t}dt\right.$$
$$\left.+\frac{1.3}{(2t)^2}\int_0^\infty\frac{t\sqrt{t}}{2!}e^{-2\lambda t}dt+\ldots\right\}.$$

The infinite integrals are known and have been evaluated. Substituting their values this series becomes:—

$$\frac{1}{\sqrt{2\pi\lambda t}}\left\{1+\frac{1}{8\lambda t}+\frac{1^2.3^2}{2!(8\lambda t)^2}+\frac{1^2.3^2.5^2}{3!(8\lambda t)^3}+\cdots\right\}$$

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]

$$I=\sqrt{\frac{C}{\pi R}}\frac{d}{dt}\int_0^t\frac{e^{-\lambda(t-\tau)}}{\sqrt{\tau}}\,d\tau$$

$$=\sqrt{\frac{C}{\pi R}}\left\{\frac{1}{\sqrt{t}}-\lambda\int_0^t\frac{e^{-\lambda\tau}}{\sqrt{t-\tau}}d\tau\right\} \tag{120}$$

by obvious transformations.

Asymptotic expansion of the definite integral as in the preceding example gives the asymptotic formula

$$I=-\sqrt{\frac{C}{\pi R t}}\left\{\frac{1}{2\lambda t}+\frac{1.3}{(2\lambda t)^2}+\frac{1.3.5}{(2\lambda t)^3}+\cdots\right\}.$$

The operational formula of the problem is

$$I=\sqrt{\frac{C}{R}}\frac{p}{p+\lambda}\sqrt{p}$$

$$=\sqrt{\frac{C}{R}}\frac{p/\lambda}{1+p/\lambda}\sqrt{p}$$

$$=\sqrt{\frac{C}{R}}\left\{\frac{p}{\lambda}-\left(\frac{p}{\lambda}\right)^2+\left(\frac{p}{\lambda}\right)^3-\cdots\right\}\sqrt{p}.$$

Applying the Heaviside Rule, we get the asymptotic expansion

$$I=-\sqrt{\frac{C}{\pi R t}}\left\{\frac{1}{2\lambda t}+\frac{1.3}{(2\lambda t)^2}+\frac{1.3.5}{(2\lambda t)^3}+\cdots\right\}$$

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,

$$I=\sqrt{\frac{C}{R}}\left\{\left(\frac{p}{\omega}\right)-\left(\frac{p}{\omega}\right)^3+\left(\frac{p}{\omega}\right)^5-\ldots\right\}\sqrt{p}$$

and explicitly

$$I=-\sqrt{\frac{C}{\pi Rt}}\left\{\frac{1}{2\omega t}-\frac{1.3.5}{(2\omega t)^3}+\ldots\right\}$$

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

$$\frac{1}{p}\phi\,(\sqrt{p})=\int_0^\infty h(t)e^{-pt}dt. \tag{123b}$$

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

$$\begin{aligned}
g(t)&=h(t)-a_0\\
g_1(t)&=g(t)-\frac{a_1}{\sqrt{\pi t}}\\
g_2(t)&=tg_1(t)+\frac{1}{2}\frac{a_3}{\sqrt{\pi t}}\\
g_3(t)&=t.g_2(t)-\frac{1.3}{2^2}\frac{a_5}{\sqrt{\pi t}}\\
g_4(t)&=t.g_3(t)+\frac{1.3.5}{2^3}\frac{a_7}{\sqrt{\pi t}}\\
&\text{- - - - - - - - - - - - - - -}
\end{aligned} \tag{123c}$$

Successive substitutions in the integral equation (123b) and repeated differentiations with respect to p, lead to the set of formulas,

$$\begin{aligned}
&\int_0^\infty g(t)e^{-pt}dt \sim \frac{a_1}{\sqrt{p}} \text{ as } p\to 0\\
&\int_0^\infty t.g_1(t)e^{-pt}dt \sim \frac{a_3}{2\sqrt{p}} \text{ as } p\to 0\\
&\int_0^\infty t.g_2(t)e^{-pt}dt \sim \frac{1.3}{2^2}\frac{a_5}{\sqrt{p}} \text{ as } p\to 0\\
&\int_0^\infty t.g_3(t)e^{-pt}dt \sim -\frac{1.3.5}{2^3}\frac{a_7}{\sqrt{p}} \text{ as } p\to 0\\
&\text{- -}
\end{aligned} \tag{123d}$$

Now *assuming* that $h(t)$ satisfies the restrictions stated in the preceding proposition, it follows from that proposition, that

$$\begin{aligned}
&g(t) \sim a_1/\sqrt{\pi t} \text{ as } t\to\infty\\
&g_1(t) \sim -\frac{a_3}{2t\sqrt{\pi t}} \text{ as } t\to\infty\\
&g_2(t) \sim \frac{1.3}{2^2 t}\frac{a_5}{\sqrt{\pi t}} \text{ as } t\to\infty\\
&g_3(t) \sim -\frac{1.3.5}{2^3 t}\frac{a_7}{\sqrt{\pi t}} \text{ as } t\to\infty\\
&\text{- - - - - - - - - - - - - - - - - -}
\end{aligned} \tag{123e}$$

From the set equations (123d) and (123e) it follows by successive substitutions that

$$h(t) \sim a_0 + \frac{1}{\sqrt{\pi t}}\left(a_1 - a_3\frac{1}{2t} + a_5\frac{1.3}{2t^2} - a_7\frac{1.3.5}{(2t)^3} + \cdots\right)$$

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

$$V=e^{-\sqrt{\alpha p}} \tag{124}$$

where $\alpha=x^2RC$, 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

$$V=1-\frac{\sqrt{\alpha p}}{1!}+\frac{\alpha p}{2!}-\frac{\alpha p\sqrt{\alpha p}}{3!}+\frac{(\alpha p)^2}{4!}-\cdots$$

which may be rearranged as

$$V=1-\left(1+\frac{\alpha p}{3!}+\frac{(\alpha p)^2}{5!}+\cdots\right)\sqrt{\alpha p}+\left(\frac{\alpha p}{2!}+\frac{(\alpha p)^2}{4!}+\frac{(\alpha p)^3}{6!}+\cdots\right) \tag{125}$$

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

$$\begin{aligned} V&=1-\left(1+\frac{\alpha}{3!}\frac{d}{dt}+\frac{\alpha^2}{5!}\frac{d^2}{dt^2}+\cdots\right)\sqrt{\frac{\alpha}{\pi t}} \\ &=1-\sqrt{\frac{\alpha}{\pi t}}\left(1-\frac{1}{3!}\left(\frac{\alpha}{2t}\right)+\frac{1.3}{5!}\left(\frac{\alpha}{2t}\right)^2-\frac{1.3.5}{7!}\left(\frac{\alpha}{2t}\right)^3+\cdots\right\} \end{aligned} \tag{126}$$

or

$$V=1-\sqrt{\frac{\alpha}{\pi t}}\left(1-\frac{1}{3}\left(\frac{\alpha}{4t}\right)+\frac{1}{5.2!}\left(\frac{\alpha}{4t}\right)^2-\frac{1}{7.3!}\left(\frac{\alpha}{4t}\right)^3+\cdots\right). \tag{127}$$

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

$$V = \sqrt{\frac{t'}{\pi}} e^{-1/t'} \left\{ 1 - \left(\frac{t'}{2}\right) + 1.3\left(\frac{t'}{2}\right)^2 - \cdots \right\}. \tag{132}$$

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_o: 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 $(1-V)/R_o$: equating these expressions we get

$$\frac{1-V}{R_o}=V\sqrt{pC/R}$$

whence

$$V=\frac{1}{\sqrt{p/\lambda}+1} \tag{133}$$

where $1/\sqrt{\lambda}=R_o\sqrt{C/R}$. 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

$$h=\frac{\sqrt{p}}{2n+1}\sum_{m=0}^{2n}\frac{p_m^{n+1}}{p-p_m}-\frac{1}{2n+1}\sum_{m=0}^{2n}\frac{p_m}{p-p_m} \tag{139}$$

where

$$p_m=e^{i\frac{2m\pi}{2n+1}} \qquad (m=0,1,2\ldots 2n).$$

Write, for convenience,

$$h=\sum_{m=0}^{2n}h_m$$

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

$$h_m(t)=\frac{1}{2n+1}\left(1+e^{p_mt}\left\{\frac{p_m^{n+1}}{\sqrt{\pi}}\int_0^t\frac{e^{p_m\tau}}{\sqrt{\tau}}d\tau-1\right\}\right) \tag{142}$$

$$=\frac{1}{2n+1}\left(\begin{array}{l}1+e^{p_mt}\left\{\dfrac{p_m^{n+1}}{\sqrt{\pi}}\displaystyle\int_0^\infty\frac{e^{-p_m\tau}}{\sqrt{\tau}}d\tau-1\right\}\\ -\dfrac{p_m^{n+1}}{\sqrt{\pi}}e^{p_mt}\displaystyle\int_t^\infty\frac{e^{-p_mt}}{\sqrt{\tau}}d\tau\end{array}\right) \tag{143}$$

$$=\frac{1}{2n+1}\left(1-\frac{p_m^{n+1}}{\sqrt{\pi}}e^{p_mt}\int_t^\infty\frac{e^{-p_m\tau}}{\sqrt{\tau}}d\tau\right). \tag{144}$$

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 (141) as

$$h_m(t) = \frac{1}{2n+1}\left(\begin{array}{l} 1-e^{p_m t} \\ \quad + \dfrac{p_m^{n+1}}{\sqrt{\pi}} \displaystyle\int_0^t \frac{e^{p_m \tau}}{\sqrt{t-\tau}}\, d\tau \end{array}\right). \tag{145}$$

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 ωt, 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:

$$\begin{aligned} x_s &= \frac{p}{\omega}\left(1+\frac{p^2}{\omega^2}\right)^{-1}\frac{1}{H(p)} \\ &= \left\{\frac{p}{\omega}-\left(\frac{p}{\omega}\right)^3+\left(\frac{p}{\omega}\right)^5-\left(\frac{p}{\omega}\right)^7+\ \ldots\right\}\frac{1}{(H(p)} \end{aligned} \tag{150}$$

and

$$\begin{aligned} x_c &= \left(\frac{p}{\omega}\right)^2\left(1+\frac{p^2}{\omega^2}\right)^{-1}\frac{1}{H(p)} \\ &= \left\{\left(\frac{p}{\omega}\right)^2-\left(\frac{p}{\omega}\right)^4+\left(\frac{p}{\omega}\right)^6-\left(\frac{p}{\omega}\right)^8+\ \ldots\right\}\frac{1}{H(p)}. \end{aligned} \tag{151}$$

Now let us identify $1/H(p)$ with $h(t)$ and replace p^n by d^n/dt^n: we get

$$x_s = \left\{\frac{1}{\omega}\frac{d}{dt}-\frac{1}{\omega^3}\frac{d^3}{dt^3}+\frac{1}{\omega^5}\frac{d^5}{dt^5}-\ \ldots\right\}h(t) \tag{152}$$

and

$$x_c = \left\{\frac{1}{\omega^2}\frac{d^2}{dt^2}-\frac{1}{\omega^4}\frac{d^4}{dt^4}+\frac{1}{\omega^6}\frac{d^6}{dt^6}-\ \ldots\right\}h(t). \tag{153}$$

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 ωt and cos ωt 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

$$x_s = \frac{d}{dt}\int_0^t \sin\,\omega\tau.h(t-\tau)d\tau = \int_0^t \sin\,\omega(t-\tau)h'(\tau)d\tau + h(o)\,\sin\,\omega t \quad (154)$$

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

$$x_s = \sin\,\omega t\int_0^\infty \cos\,\omega t.h'(t)dt - \cos\,\omega t\int_0^\infty \sin\,\omega t.h'(t)dt$$

$$+h(o)\sin\,\omega t - \int_t^\infty \sin\,\omega(t-\tau)h'(\tau)d\tau. \quad (155)$$

The first three terms are simply the steady-state response to the impressed e.m.f. sin ωt: 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

$$T_s = -\int_t^\infty \sin\,\omega(t-\tau)h'(\tau)d\tau. \quad (156)$$

Now from (156)

$$T_s = -\frac{1}{\omega}\int_t^\infty h'(\tau).d.\cos\omega(\tau-t)$$

and integrating by parts

$$T_s = \frac{1}{\omega}\frac{d}{dt}h(t) + \frac{1}{\omega}\int_t^\infty \cos\,\omega(\tau-t)\frac{d^2}{dt^2}h(\tau)d\tau. \quad (157)$$

Repeating the process of partial integration, we get:

$$T_s = \frac{1}{\omega}\frac{d}{dt}h(t) - \frac{1}{\omega^2}\int_t^\infty \sin\,\omega(\tau-t)\frac{d^3}{d\tau^3}h(\tau)d\tau. \quad (158)$$

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

$$T_s=\left(\frac{1}{\omega}\frac{d}{dt}-\frac{1}{\omega^3}\frac{d^3}{dt^3}+\frac{1}{\omega^5}\frac{d^5}{dt^5}-\cdots+\frac{(-1)^{n-1}}{\omega^{2n-1}}\frac{d^{2n-1}}{dt^{2n-1}}\right)h(t)$$
$$+\frac{(-1)^n}{\omega^{2n}}\int_t^\infty \sin\,\omega(\tau-t)\,\frac{d^{2n+1}}{dt^{2n+1}}h(\tau)d\tau. \qquad (159)$$

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

$$\left(\frac{1}{\omega}\frac{d}{dt}-\frac{1}{\omega^3}\frac{d^3}{dt^3}+\frac{1}{\omega^5}\frac{d^5}{dt^5}-\cdots+\frac{(-1)^{n-1}}{\omega^{2n-1}}\frac{d^{2n-1}}{dt^{2n-1}}\right)h(t) \qquad (160)$$

with a remainder term

$$\frac{(-1)^n}{\omega^{2n}}\int_t^\infty \sin\,\omega(\tau-t)\frac{d^{2n+1}}{dt^{2n+1}}\,h(\tau)d\tau.$$

If the impressed e.m.f. is $\cos \omega t$, the corresponding series for the transient distortion, T_c, is

$$\left(\frac{1}{\omega^2}\frac{d^2}{dt^2}-\frac{1}{\omega_4}\frac{d_4}{dt_4}+\frac{1}{\omega^6}\frac{d^6}{dt^6}-\cdots-\frac{(-1)^n}{\omega^{2n}}\frac{d^{2n}}{dt^{2n}}\right)h(t) \qquad (161)$$

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 2475°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.[8] 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.

[8] 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.