Bound Periodical

528749

Kansas City Public Library

THE SCHOOL DISTRICT OF KANSAS CITY, MISSOURI

This Volume is for

REFERENCE USE ONLY

11-26—6m—P

PUBLIC LIBRARY
KANSAS CITY
MO

SEP 6 '4

From the collection of the



San Francisco, California
2008

PUBLIC LIBRARY
KANSAS CITY
MO

PUBLIC LIBRARY
KANSAS CITY
MO.

THE BELL SYSTEM TECHNICAL JOURNAL

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

EDITORIAL BOARD

J. J. Carty Bancroft Gherardi F. B. Jewett
E. B. Craft L. F. Morehouse O. B. Blackwell
H. P. Charlesworth E. H. Colpitts H. D. Arnold
R. W. King—*Editor* J. O. Perrine—*Asst. Editor*

VOLUME V
1926

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

v.5
1926

Bound
Periodical

528749

Ap 15 '27



JOSEPH HENRY

1799–1878

The Bell System Technical Journal

January, 1926

Joseph Henry

The American Pioneer in Electrical Communication

By BANCROFT GHERARDI and ROBERT W. KING

IN the brilliant galaxy of investigators to whom we owe our knowledge of electrical science, Joseph Henry stands out as of the first magnitude; and for those who are associated with the Bell System, the present is a most appropriate time to review his researches which had an important guiding influence on the development of electrical communication. The present year marks the fiftieth since the invention of the telephone by Alexander Graham Bell, and among the scientists with whom Bell conferred at that time, he gave a place of honor to Henry. In a letter to his parents written in March, 1875, while he was busy in an effort to perfect the harmonic telegraph, and before he had turned his attention to the telephone, Bell wrote:

"Now to resume telegraphy. When I was in Washington, I had a letter of introduction to Professor Henry, who is the Tyndall of America. I had found on inquiry at the Institute of Technology, that some of the points I had discovered in relation to the application of acoustics to telegraphy had been previously discovered by him. I thought I would, therefore, explain all the experiments, and ascertain what was new and what was old. He listened with an unmoved countenance, but with evident interest to all, but when I related an experiment that at first sight seems unimportant, I was startled at the sudden interest manifested.

"I told him that on passing an intermittent current of electricity through an empty helix of insulated copper wire, a noise could be heard proceeding from the coil, similar to that heard from the telephone. He started up, said, 'Is that so? Will you allow me, Mr. Bell, to repeat your experiments, and publish them to the world through the Smithsonian Institute, of course, giving you the credit of the discoveries?'

"I said it would give me extreme pleasure, and added that I had apparatus in Washington, and could show him the experiments myself at any time. . . .

"We appointed noon next day for the experiments, I set the in-

strument working and he sat at a table for a long time with the empty coil of wire against his ear listening to the sound. I felt so much encouraged by his interest that I determined to ask his advice about the apparatus I have designed for the transmission of the human voice by telegraph. I explained the idea and said, 'What would you advise me to do, publish it and let others work it out, or attempt to solve the problem myself?' He said he thought it was 'the germ of a great invention,' and advised me to work at it myself instead of publishing. I said that I recognized the fact that there were mechanical difficulties in the way that rendered the plan impracticable at the present time. I added that I felt that I had not the electrical knowledge necessary to overcome the difficulties. His laconic answer was, 'GET IT.'

"I cannot tell you how much these two words have encouraged me. Such a chimerical idea as telegraphing vocal sounds would indeed to most minds seem scarcely feasible enough to spend time in working over. I believe, however, that it is feasible, and that I have got the cue to the solution of the problem.

"Professor Henry seemed to be much interested in what I told him, and cross-questioned me about my past life, and specially wanted to know where I had studied physics . . . . "

Joseph Henry was born in Albany, New York, in 1799, and coming to full maturity of mind at the beginning of a century which will probably never be surpassed for fruitful research in the field of electricity, he demonstrated, at the very outset of his career, his right to stand for all time with the foremost investigators in this department of natural science. Henry was, moreover, a many-sided man. His distinguished career leads into many fields and before reviewing his researches on electro-magnetism we may note briefly the very diversified and yet important character of his other work.

During the latter half of his life, official duties as the director of the Smithsonian Institution consumed an ever increasing portion of his time, but he still found opportunity to prosecute many original inquiries,—for example, into the application of acoustics to building, into the best construction and arrangement of lecture rooms, and into the strength of various building materials. As one of his first administrative acts, he organized a widespread corps of observers for simultaneous weather and meteorological reports by means of the telegraph which was yet in its infancy. He was the first to have the daily atmospheric conditions indicated upon a map of the country and to utilize this information in making weather forecasts.

He was an active and long-standing member of the Lighthouse Board of this country and his diligent investigations into the efficiency of various illuminants and the best conditions for their use greatly improved the beacons which dotted our coasts. During the dark days of the Civil War, Henry clearly saw the tremendous advantage to be derived from a mobilization of the nation's scientific men for cooperative service. His vision, backed by his tremendous energy and ability, resulted in the formation of the National Academy of Sciences, under a Congressional charter signed by Abraham Lincoln.

More than fifty years later this same National Academy of Sciences was again called upon in time of national need, and, using the mechan-

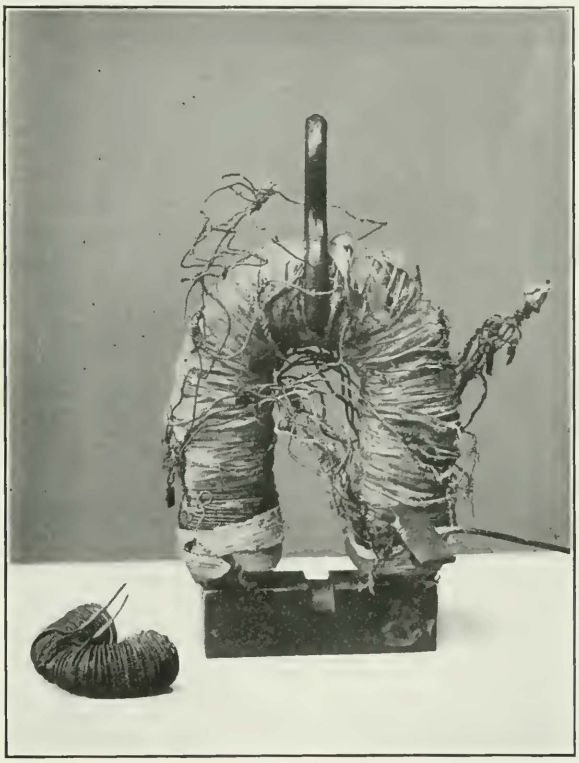


Fig. 1

ism inspired by Henry, there re-occurred, in 1916, under presidential proclamation, a mobilization of the nation's scientific and technical men.

While the details of Henry's life and work are perhaps not widely

known, his researches are of the most enduring character, and for all time must enter intimately into the lives of all civilized mankind. He was without peer among the American physicists of his time, and it is well attested by every record that he was a man of varied culture, of large breadth and liberality of views, of generous impulses, of great gentleness and courtesy of manner, combined with equal firmness of purpose and energy of action.

Let us now turn to Henry's investigations of electro-magnetism, which were among his earliest scientific undertakings. He began his career in 1826 in New York State at the Albany Academy, where he had only the apparatus he could construct with his own hands and, out of each year, but a single month uninterrupted by other duties to devote to his researches. It was there—independently of Faraday and on some fundamental points prior to him—that Henry discovered the laws of current induction. At the same time he undertook a study of the electromagnet which prepared the way for not only the telephone and telegraph, but also for all types of dynamos and motors.

The electromagnet was discovered by Sturgeon in England, but Henry's contributions to our knowledge of it were so great that after his work, a powerful instrument suitable for many uses replaced what had been a feeble toy. When he started his work on the electromagnet its design was not understood; when he had completed his work he had developed a magnet, the design of which was understood and which could be adapted, according to the rules which he laid down, to a multitude of purposes.

With reference to the making of electromagnets, Henry pointed out the improvements which resulted from insulating the conducting wire itself, instead of the rod to be magnetized, and by covering the whole surface of the iron with a series of coils in close contact. This was effected by insulating a long wire with silk thread, and winding this around the rod of iron in close coils from one end to the other. The same principle was extended by employing a still longer insulated wire, and winding several strata of this over the first, care being taken to insure the insulation between each stratum by a covering of silk ribbon. By this arrangement the rod was surrounded by a compound helix formed of a long wire of many turns instead of a single helix of a few turns.

Thus Henry laid down the rules, which, in general, are followed today in the construction of commercial electromagnets; namely, that the wire should be insulated, that it should be wound in layers, and that there should be several layers, one above the other. He

also did another thing in his actual construction: he adopted what may be called the spool construction, the placing of the windings on spools, and then the sliding of the spools on the core. That is a standard method of building electromagnets today.

Soon after doing this work Henry built a magnet to be used at Yale University, which was in its time a wonder and would even today be considered very powerful. He also built a series of magnets in which the emphasis was placed upon the lifting power in relation to the weight of the magnet and succeeded in designing one which, when energized by a single small cell, could support 420 times its own weight.

The improvements which Henry made in magnets suggested to him applications of magnetic attraction to the production of mechanical motion. He realized that electromagnets such as he built were easy to control, and believed that he could design a machine by which he could get power from an electric current and this at a time when the only source of current were primary batteries as the dynamo did not yet exist.

His electric motor was the first ever built to use electromagnets;[1] it was extremely simple consisting of an electromagnet supported at its center by a pivot so that it could rock back and forth under the alternating pulls of two permanent magnets. To effect the reversal of magnetization of the electromagnet and hence the alternation of pulls, mercury cups were arranged so that wires would dip in them as the suspended magnet rocked to and fro. These contacts were the prototype of the commutator which is found in every direct current motor and dynamo today. It is interesting to note the words in which Henry described this invention. In Silliman's *American Journal of Science* for 1831 he wrote, "I have lately succeeded in producing motion in a little machine by a power which I believe has never before been applied in mechanics—by magnetic attraction and repulsion. Not much importance, however, is attached to the invention since the article in its present state can only be considered a philosophical toy; although in the progress of discovery and invention it is not impossible that the principle or some modification of it on a more extended scale may hereafter be applied to some useful purpose."

The modesty of this statement and Henry's vision of the future possible applications of the principle there shown cannot fail to com-

[1] Faraday has some years before shown that a wire carrying a current could be caused to revolve continuously around the pole of a permanent magnet. Henry's advance over this was considerable in that he materially increased the force causing motion by employing the attraction between two magnets, one permanent and one generated by current. The motor using electromagnets throughout did not come until later.

mand our admiration. Of course, until the dynamo was invented at a later date, and a substantial electric current became available, the motor could not be much more than he characterized it, "a philosophical toy."

Henry also became interested in a determining whether an electromagnet could be operated from a distance so that the doing of some work—for example the ringing of a bell—could be controlled from a distant station. From his investigations directed to this end, Henry was the first to appreciate that the effect of the resistance of long lengths of wire to the passage of electric current could be minimized by properly proportioning the battery and the magnet windings to the length and resistance of the line wires.

Efforts had been made by others prior to Henry's time to devise successful electric telegraphs. They had failed, however, because they did not know how to proportion their magnets and their batteries so as to operate over any substantial length of line. The literature of that time contains a number of demonstrations of the impossibility of operating an electric telegraph, because scientists could arrange instruments which would operate successfully when separated by a few feet, or even one hundred feet, but they would not work at a distance of thousands of feet because of the resistance of the long line wire.

What Henry did was to determine the proportioning of the various parts of the system so as to secure operation. He found, when his magnet was connected by a short wire to the battery, that the greatest magnetizing effect was obtained by joining the cells of the battery in parallel, but that a series arrangement of the battery would give the greatest pull if a long wire (a length of a mile or more was used in some of his experiments) carried the current. He also obtained the best operation over a short line when the magnet winding consisted of several distinct coils, all connected in multiple; and for operation over a long line he found it best either to connect these coils in series or to apply to the magnet a single long winding. Henry was therefore the first to produce an electric telegraph, and more than that, the transmission of electrical energy to a distance. That first telegraph paved the way for all the telegraph systems, all the ocean cable systems, and contained the principle of all telephone call bells.

One of Henry's greatest discoveries from the standpoint of electrical science, but a discovery in which he must yield the first place to Faraday, is that of mutual induction—the fact that a wire when moving with respect to a magnetic field has an electromotive force generated in it. Although Henry made his discovery independently

of Faraday, the latter was the first to make known his observations to the world, and it is no trifling index of Henry's character that he never in any way intimated that he was entitled to share with Faraday credit for the discovery.

Because Henry was anticipated in the publication of his observation of mutual induction, he does not appear to have left a verbal record of the steps of reasoning by which he was led to the discovery. However, he does tell us what the arrangement of apparatus was and if we bear in mind that he was seeking a method of generating an electric current from a magnet—this magnet, in turn, being itself the product

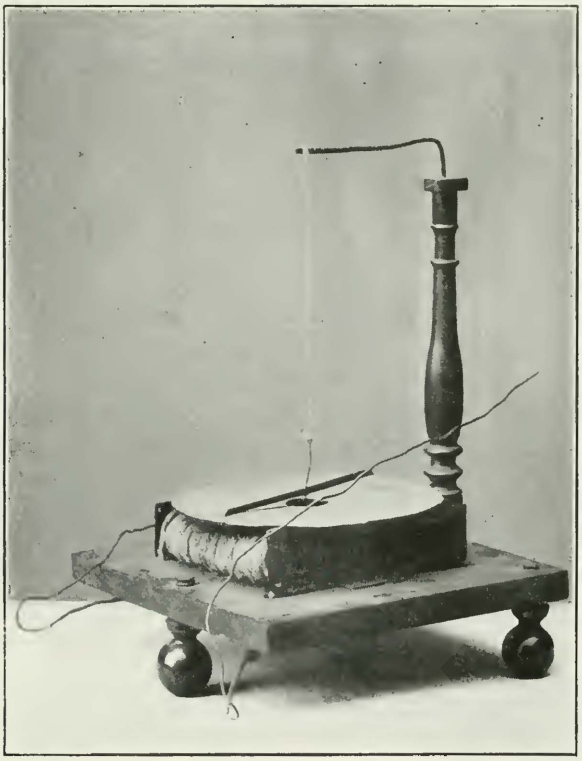


Fig. 2

of a current—we cannot but be impressed by the directness of his method.

Writing of his original observations, Henry says he "succeeded in producing electrical effects in the following manner, which differs

from that employed by Mr. Faraday and which appears to me to develop some new and interesting facts. A piece of copper wire, about thirty feet long and covered with elastic varnish, was closely coiled around the middle of the soft iron armature of a galvanic magnet . . . which, when excited will readily sustain between six and seven hundred pounds. The armature thus furnished with wire was placed in its proper position across the ends of the magnet and fastened so that no motion could take place. The two projecting ends of the helix were connected with a distant galvanometer by means of two copper wires each about forty feet long. This arrangement being completed, I stationed myself near the galvanometer and directed an assistant at a given word to suddenly immerse the galvanic battery attached to the magnet. At the instant of immersion the north end of the needle was deflected 30° to the west, indicating a current of electricity from the helix surrounding the armature. The effect, however, appeared only as a single impulse, for the needle after a few oscillations, resumed its former undisturbed position, although the action of the battery was still continued. I was, however, much surprised to see the needle suddenly deflected from a state of rest to about 20° to the east, when the battery was suddenly withdrawn from the acid, and again deflected to the west when it was re-immersed. This operation was repeated many times in succession, and uniformly with the same result."

It was in this same paper that Henry announced his observation of the phenomenon of self-induction, a most important discovery and one for which he holds full credit for having first made it known to the world. He writes, "I may, however, mention one fact which I have not seen noticed in any work, and which appears to me to belong to the same class of phenomena as those before described; it is this: when a small battery is moderately excited by diluted acid, and its poles, which should be terminated by cups of mercury, are connected by a copper wire not more than a foot in length, no spark is perceived when the connection is either formed or broken; but if a wire of thirty or forty feet long be used instead of the short wire, though no spark will be perceptible when the connection is made, yet when it is broken by drawing one end of the wire from its cup of mercury, a vivid spark is produced . . . . . The effect appears somewhat increased by coiling the wire into a helix." In a somewhat later paper we find the following statement. "A ribbon of sheet copper nearly an inch wide, and twenty-eight and a half feet long, was covered with silk, and rolled into a flat spiral similar to the form in which woolen binding is found in commerce. With this a

vivid spark was produced, accompanied by a loud snap. The same ribbon uncoiled gave a feeble spark."

Henry tried many modifications of this experiment and in the end drew the conclusion that the after-current he was observing was due to the inductive effect of the current in the wire upon itself, and that this became particularly apparent when the wire was so coiled that its various turns lay close together. The discovery of mutual induction by Faraday and the discovery of self-induction by Henry constitute two halves of a whole, and it is appropriate that to these men should go equal recognition in the matter of having electrical units names after them. Of the three units by which the properties of every electric circuit are measured, the unit of capacity was named after Faraday, and unit of inductance after Henry; the third unit, that of resistance, recognizes the fundamental researches of Ohm.

A few years later, after having accepted the chair of physics at Princeton University, Henry returned to the subject of induced currents. In his earlier work he, like Faraday, had used the continuous currents which a voltaic battery generates. He now chose the currents which flow when a Leyden jar is discharged. To register the inductive effects of the fleeting currents of discharge Henry adopted a device consisting of an unmagnetized needle placed in a small coil of wire. Through this coil the induced current had to flow. The use of the needle as an indicator led Henry to an important observation. He noticed that following a discharge, the direction of magnetization of the needle depended upon the distance across which the inductive effect had occurred. To account for this curious result, he advanced the hypothesis—later shown to be correct—that the discharge is oscillatory.

Here was the germ of a great discovery. The oscillatory character of the discharge is one of the fundamental and important properties of certain types of electric circuit. Henry did not have the facilities, however, for carrying his investigations in this field far enough to attract the attention of the scientific world. It was not until 1855, some thirteen years later, when Lord Kelvin was led independently by mathematical considerations to believe that the discharge is oscillatory, that the significance of the phenomenon began to be understood.

Henry's work contained the germ of yet another important discovery. Some of his experiments on induction by Leyden jar discharges involved the transmission of electric force without wires through distances as great as two hundred feet, and through the floors and walls of buildings. And in similar experiments in which he

observed the effects of lightning flashes in place of sparks from a Leyden jar, he found that he could get the lightning to magnetize needles up to a distance as great as eight miles. This was about 1842. Here we have the earliest evidence of ether waves of the type that the radio engineer employs. But again the significance of Henry's work was not recognized. This could only have come after much fuller investigation. However, it is instructive to reflect for a moment on what might have been had Henry possessed the time and facilities for carrying his work further. Needless to say, there is a wide gulf between the wireless telegraph of today and its earliest precursor with which Henry received an electromagnetic signal from a lightning flash eight miles away, but it is wholly possible that, had Henry not been called to other work, the world might have possessed a wireless telegraph capable of sending messages over substantial distances many years before it did.

Writing of Henry, Simon Newcomb, the celebrated astronomer said,[2] "His scientific work is marked by acuteness in cross-examining nature, a clear appreciation of the logic of science, and an enthusiasm for truth without respect to its utilitarian results." A man of the highest scientific ability, Henry spent the better part of his life as the head of an institution dedicated to "the increase and diffusion of knowledge among men."

"The mantle of Franklin has fallen upon the shoulders of Henry," wrote Sir David Brewster,[3] the eminent English scientist, and it is reported that Abraham Lincoln declared, when he became acquainted with Henry after assuming the Presidency, "The Smithsonian Institution must be a grand school if it produces such thinkers as Henry." He was, in every way and in the best that the word implies, a scientist, and the interest in scientific questions which dominated his life, remained with him to the very end,—almost the last words to pass his lips were whether the transit of the planet mercury had been successfully observed. If we use the word "Dean"—so rich in academic association—to stand at once for the greatest usefulness to one's fellowmen as well as for the highest achievements in the field of scholarship and research, for lifelong devotion to public service, for breadth of view and tolerance regarding all questions, whether arising in science or directly out of human relations, and as epitomizing all that is best and highest in man's intellectual life, we may well call Joseph Henry the Dean of American scientists.

[2] Biographical Memoir; National Academy of Sciences, Apr. 21, 1880.
[3] Biographical Memoir; prepared by Prof. Asa Gray in behalf of the Board of Regents of the Smithsonian Institution.

Correction of Data for Errors of Measurement

By W. A. SHEWHART

Introduction

EVERY measurement is subject to error. This universally accepted truth is the result of every-day experience. From the simplest type of measurement, such as determining the length of a board with an ordinary tape measure, to the most refined type of measurement, such as determining the charge on an electron, errors are bound to creep in.

Now, a manufacturer must constantly make measurements of one kind or another in an effort to control his production processes and to measure the quality of his finished product in terms of certain of its characteristics, but, before he can safely determine the significance of observed differences in his production processes or in the quality of his product as given by these measurements, he must make allowance for his errors of measurement; i.e., for the fact that the observed differences may be larger or smaller than the true differences. To make such allowances for the errors of measurement of any characteristic, to find out what the true magnitude of the characteristic most probably is, to find out, as it were, what a thing most probably is from what it appears to be, presents an endless chain of interesting problems to be solved.

Three important types of problems arising in engineering practice are discussed in this paper. They are:

1. Error correction of data taken to show the quality of a particular lot.

2. Error correction of data taken periodically to detect significant changes in quality of product.

3. Error correction of data taken to relate observed deviations in quality of product to some particular cause.

The solution of the first one is presented here for the first time. The solution of the second has been generalized to include cases not previously solvable. All three types of problems are illustrated.

PART I

Type 1—Error Correction of Data Taken to Show the Quality of a Particular Lot

Let us take a specific problem first. Assume that we have a lot consisting of 15,000 transmitters [1] and a machine with which to measure the efficiency of each instrument. Suppose we make one observation on each transmitter—a total of 15,000 measurements. Suppose we find, as in the distribution illustrated in Fig. 1, that one measurement is in the efficiency range −1.75 to −1.50, 17 within the range

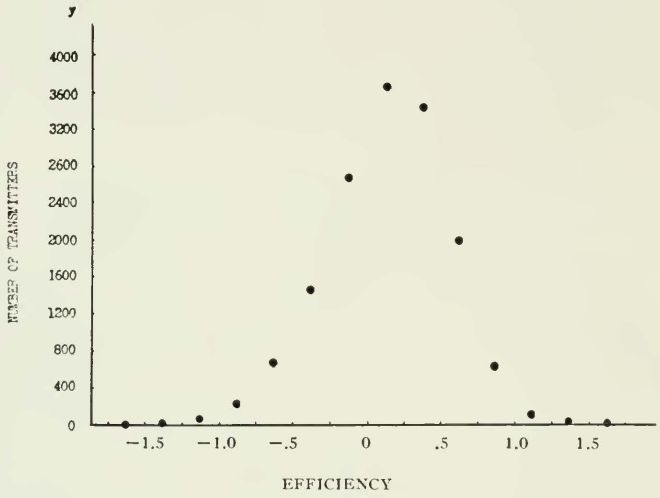


Fig. 1—Typical frequency distribution. Chart showing observed number of transmitters versus efficiency

−1.50 to −1.25 units, and so on. The vertical height of a point represents the number or frequency of occurrence of observations falling within the corresponding interval laid off on the horizontal axis of the chart.

So far so good, but suppose a customer wants to buy these transmitters. We know that some transmitter which appeared to have an efficiency within the range of 1.25 to 1.50 units say, *may* actually have had an efficiency within some other interval. We know too that, because of the errors of measurement, the transmitters appear to differ more among themselves than they really do. We therefore

[1] Of course, the efficiency of a transmitter does not remain constant during a series of tests but these inherent variations in the transmitter may be considered, for our purpose, as forming a component part of the resultant error of measurement.

desire to find the most probable numbers of transmitters within the different intervals indicated in Fig. 1.

Analytical Statement of Problem

Let us assume that the most probable number of transmitters within the interval of efficiency from X to $X+dX$ is $f_T(X)dX$. It is this function $f_T(X)$ that we want to find. Similarly let us assume that there is some function $f_o(X)$ such that $f_o(X)dX$ gives the observed number of transmitters appearing to have efficiencies within the interval X to $X+dX$ where the measurements are made by a method wherein the probability of making an error within the interval x to $x+dx$ is $f_E(x)dx$. It is reasonable to expect that, if two of these functions are known, the third can be easily determined. We shall proceed to show that this is the case. Let us first find the law of error experimentally.

Finding the Law of Error

The problem is to determine the chance of making an error of a given magnitude in measuring the efficiency of any transmitter. Naturally, the only way of doing this is to make a series of measurements on a single transmitter from which we can determine the observed frequency of occurrence of measurements which differ from the average by some fixed amount, and thus find what percentage of the total number of measurements may be expected to fall within any given range on either side of the average. Common sense and intuition may tell us that we may expect to find a large percentage of the measurements within a narrow range on either side of the average, that there will be just as many measurements greater than the average by a certain amount as there are less than the average by the same amount, and that large deviations from the average may be expected to occur with less frequency than small deviations. Suppose we make 500 observations of the efficiency of a single transmitter and find the distribution given in Fig. 2. Just as we might have expected, the observed values of the efficiency of the transmitter are grouped symmetrically about the average of all the observed values. We see that the maximum deviation between observations on a single transmitter is quite large (33%) compared with the actual maximum differences observed between the efficiencies of the transmitters.

The results reproduced in Fig. 2 suggest that the deviations for the case in hand are distributed in a manner closely approximating the

bell-shaped distribution so familiar in the theory of errors. We often find, as we do in this case, that the observed distribution can be closely approximated by a function $f_E(x)$ of the form

$$f_E(x)dx = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\overline{X})^2}{2\sigma^2}} dx, \tag{1}$$

where $f_E(x)dx$ is the probability that an error x will lie within the interval x to $x+dx$, σ is the root mean square or standard deviation, $\overline{X}$ is the arithmetic mean value and $(X-\overline{X})$ is the deviation x. The



Fig. 2—Typical form of distribution of errors of measurement. Chart showing number of measurements on a single transmitter versus efficiency

function $f_E(x)$ is referred to in the literature as the normal law of error. If we try to fit such a curve to the deviations[2] given in Fig. 2, we obtain the results shown in Fig. 3. This figure is the same as Fig. 2 except for the addition of the smooth normal curve of error calculated for the observed data. Without further consideration, we shall assume the law of error to be normal and hence of the form indicated by Equation (1).

Finding the True Distribution $f_T(X)$

We have next to consider the choice of the function to represent the true distribution $f_T(X)$. Often we have reason to believe that this

[2] If the average of the observed values of the 500 observations of efficiency given in Fig. 3 is assumed to be the true value of the efficiency of the transmitter, then the deviation of an observed value from this mean is also the error of this observed value. We shall use the terms "error" and "deviation" interchangeably in this sense.

is also approximately normal, and hence we shall consider first the method for finding the observed distribution $f_o(X)$ for the special case when both the true distribution $f_T(X)$ and the law of error $f_E(X)$ are normal; i.e., when they are both of the form given by Equation (1).

We shall first obtain an experimental answer to this problem. Suppose we take, say, 1,000 instruments of some kind which are

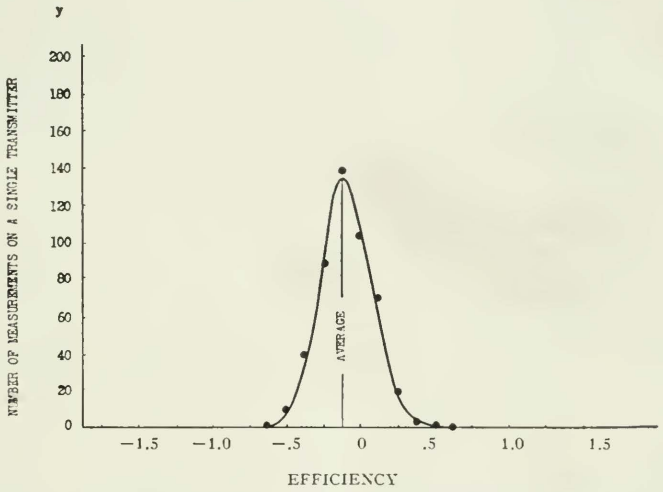


Fig. 3—Chart showing the observed distribution of errors fitted by a typical smooth curve. Data of Fig. 2 fitted by normal law of error, Eq. 1

known to be distributed in normal fashion, in respect to some characteristic, with a standard deviation σ_T. Let us measure each of these instruments by a method subject to the normal law of error whose standard deviation σ_E is $\frac{1}{2}\ \sigma_T$. The results of one such experiment are given in Fig. 4. The observed frequencies of occurrence are represented by the circles. It was found that this observed distribution could be closely approximated by a normal law $f_o(X)$ for which the standard deviation σ_o was $\sqrt{\sigma_T^2+\sigma_E^2}$. This experiment suggests a general theorem which will be demonstrated analytically in a succeeding paragraph. The theorem is: When the true distribution $f_T(X)$ and the law of error $f_E(x)$ are both normal (hence expressible in form indicated by Equation (1)) with root mean square or standard deviations σ_T and σ_E respectively, the most probable observed distribution will be normal in form with a standard deviation $\sigma_o=\sqrt{\sigma_T^2+\sigma_E^2}$.

The observed distribution in Fig. 1 is asymmetrical and hence not

normal as it should be if $f_T(X)$ and $f_E(x)$ were both normal. We must therefore, try some other function for $f_T(X)$.

Of course, experiments might be performed for other types of true and error distributions, but in all such cases the results, as in the illustration just considered, would be subject to errors of sampling.

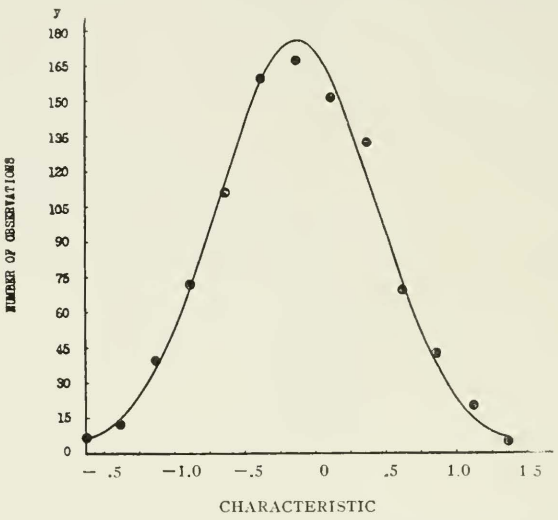


Fig. 4—Experimental results shpwing effects of errors of measurement. Normal curve fitted to observed points, when the true distribution and the law of error are both normal

Hence we shall proceed at once to the analytical treatment of the problem.

Assuming the law of error to be normal, we see that the fraction $f_E(x)dx$ of the number of objects having magnitudes between $X+x$ and $X+x+dx$ will be measured with an error between $-x$ and $-x-dx$ and hence will be observed as of magnitude X (Fig. 5). Thus

$$f_o(X) = \int_{-\infty}^{\infty} f_T(X+x)\frac{1}{\sigma_E\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma_E^2}}dx. \tag{2}$$

For the particular case treated in a previous paragraph where both the true distribution $f_T(X)$ and the law of error $f_E(x)$ are normal, we may write Equation (2) in the form

$$f_o(X)dX = \frac{1}{\sigma_T\sigma_E 2\pi}\int_{-\infty}^{\infty} e^{-\frac{(X+x)^2}{2\sigma_T^2}}e^{-\frac{x^2}{2\sigma_E^2}}dX\,dx \tag{3}$$

where σ_T and σ_E are the root mean square or standard deviations of

the true and error distributions respectively. Integration of Equation (3) gives [3]

$$f_o(X) = \frac{1}{\sigma_o \sqrt{2\pi}} e^{-\frac{X^2}{2\sigma_o^2}}, \tag{4}$$

where

$$\sigma_o = \sqrt{\sigma_T^2 + \sigma_E^2}. \tag{5}$$

Equations (4) and (5) are the analytical expression for the rule stated previously, for finding the observed distribution $f_o(X)$ when both the true and error distributions are normal, because Equation (4)

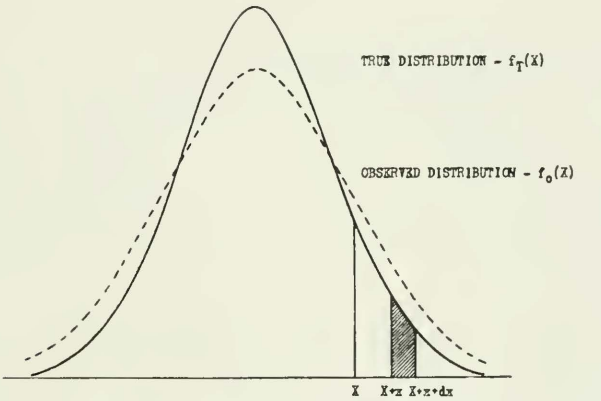


Fig. 5—Chart used in explaining the derivation of $f_o(X)$ in terms of $f_T(X)$

shows it to be normal and Equation (5) expresses the standard deviation σ_o of the observed values in terms of those of the true values and of the errors.

In practice, however, we often find that the true distribution is non-symmetrical or skew and can be more nearly approximated by the function [4]

$$f_T(X) = \frac{1}{\sigma_T \sqrt{2\pi}} e^{-\frac{X}{2\sigma_T^2}} \left[1 - \frac{k_T}{2} \left(\frac{X}{\sigma_T} - \frac{X^3}{3\sigma_T^3} \right) \right] \tag{6}$$

where k_T is a measure of the asymmetry or skewness, the modal or most probable value of X being at a distance $-\frac{k\sigma_T}{2}$ from the average

[3] See Appendix 1 where another method of solution is given.

[4] This is often referred to in the literature of statistics as the second approximation. It is in fact the first two terms of the Gram-Charlier series.

value of X. Substitution of this expression and a normal error function in Equation (2), yields upon integration [5] the following distribution $f_o(X)$ of the observed values

$$f_o(X) = \frac{1}{\sigma_o\sqrt{2\pi}} e^{-\frac{X}{2\sigma_o^2}} \left[1 - \frac{k_o}{2} \left(\frac{X}{\sigma_o} - \frac{X^3}{3\sigma_o^3} \right) \right] \tag{7}$$

where

$$\sigma_o = \sqrt{\sigma_T^2 + \sigma_E^2}, \tag{5}$$

and

$$k_o = k_T \frac{\sigma_T^3}{\sigma_o^3}. \tag{8}$$

We see that the distribution $f_o(X)$, Equation (7), of the observed values is of the same form as that $f_T(X)$, Equation (6), of the true values. The standard deviation of the errors of measurement σ_E, as in the previous case, has equal weight with the standard deviation σ_T in influencing the standard deviation σ_o of the observed values. The degree of asymmetry of the observed distribution as measured by the skewness k_o is, however, less (Equation (8)) than that of the true distribution as measured by the skewness k_T of the true distribution.

Now we can correct the observed distribution, Fig. 1, for the errors of measurement, because we find that the observed frequencies,Fig. 1, can be closely approximated by a function of the type defined by Equation (7). Knowing that the law of error, Fig. 3, is normal we conclude that the true distribution $f_T(X)$ must be a function of the same type as $f_o(X)$ was found to be except that the true standard deviation σ_T will be, from Equation (5), $\sqrt{\sigma_o^2 + \sigma_E^2}$ and the true skewness k_T will be, from Equation (8), $\frac{\sigma_o^3}{\sigma_T^3} k_o$. Now, σ_o and k_o can be calculated from the observed distribution, Fig 1, and σ_E can be determined by the data given in Fig. 3.

Thus finding the values of σ_T and k_T and substituting them in Equation (6), we have the function $f_T(X)$ representing the true distribution which we started out to find. From this knowledge of $f_T(X)$ we can now get the most probable frequencies of occurrence of the different efficiencies. Subtracting these frequencies from those observed and shown in Fig. 1, we get the corrections plotted in Fig. 6, expressed as percentages of the observed frequencies.

[5] This solution is also obtained by another method in Appendix 1.

Summary

We are now in a position to summarize the practical routine to be followed in finding the most probable distribution $f_T(X)$ of quality when the observed distribution is given.

To find $f_T(X)$, we must first know the law of error $f_E(x)$. We must show this to be normal and find the standard deviation σ_E

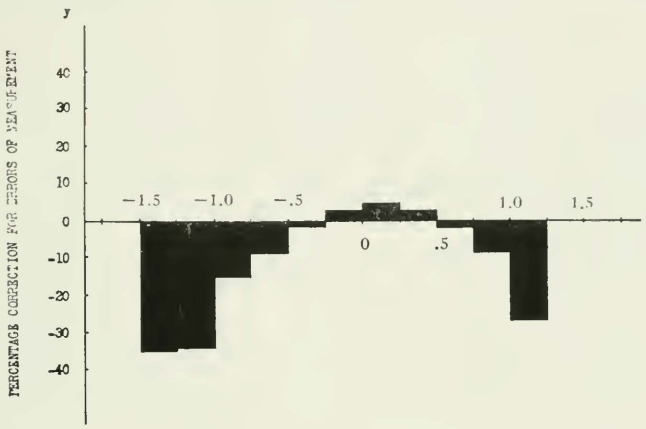


Fig. 6—Correction which must be applied to the observed distribution of transmitters Fig. **1,** because of the existence of errors of measurement

by multiple tests on a single unit. The error made in determining the standard deviation σ_E from n observations is $\frac{\sigma_E}{\sqrt{2n}}$. Hence the precision we attain in finding $f_T(X)$ depends upon the number of observations n made in finding σ_E.

Having found σ_E to the required degree of precision, we must next discover whether or not the observed distribution $f_T(X)$ is either normal or the second approximation. Standard statistical methods can be used for this purpose.

If the $f_o(X)$ is normal, we then know that

$$f_T(X) = \frac{1}{\sqrt{2\pi(\sigma_o^2 - \sigma_E^2)}}\, e^{-\frac{X^2}{2(\sigma_o^2-\sigma_E^2)}},$$

and, if $f(X)$ is second approximation, we know that $f_T(X)$ is given by Equation (6), where σ_T and k_T can be found with the aid of Equa-

tions (5) and (8) in terms of the observed values of σ_E, σ_o and k_o. In other words we have

$$f_T(X) = \frac{1}{\sqrt{2\pi(\sigma_o^2 - \sigma_E^2)}} e^{-\frac{X}{2(\sigma_o^2 - \sigma_E^2)}} \left[1 - \frac{k_o \sigma_o^3}{(\sigma_o^2 - \sigma_E^2)^{\frac{3}{2}}} \left(\frac{X}{(\sigma_o^2 - \sigma_E^2)^{\frac{1}{2}}} - \frac{X^3}{3(\sigma_o^2 - \sigma_E^2)^{\frac{3}{2}}} \right) \right].$$

PART II

Correction of Data Taken Periodically to Detect Significant Changes in Quality of Product

Irrespective of the care taken in defining and controlling the manufacturing processes, the units of a product will differ among themselves in respect to any measurable characteristic. Random fluctuations in such factors as humidity, temperature, grade of raw material, and wear and tear on machinery may produce such differences between units of a product. Such random variations in the factors underlying the manufacturing process usually yield a product in which the units differ in random fashion according to some law of probability.

Customarily, product is inspected periodically, and the data are analyzed to determine if the observed difference in two samples is greater than can be accounted for as a random variation. If it is, we may assume that the manufacturing processes have changed significantly for some reason which further investigation should disclose. Now, the presence of errors of measurement effectively increases the magnitude of the random differences to be expected from one sample to another and hence makes it harder for us to detect trends or fluctuations in product. Let us investigate this effect of errors of measurement.

Symbolic Statement of Problem

Symbolically we may assume that the probability of production of a unit of product having a characteristic X within any range X to $X + dX$ is $f_T(X)dX$, where the characteristic X is measured by a method subject to a law of error $f_E(x)$, so that $f_E(x)dx$ represents the probability of occurrence of an error x within the range x to $x+dx$. The problem is to find the corresponding distribution $f_o(X)$ for the observed magnitudes.

General Solution of Problem

Obviously the observed magnitude X_o is the algebraic sum of the true value X and the error x. Assuming that there is no correlation between these two quantities, the probability of a unit having a value of X within the range X to $X + dX$ being measured with an error x within the range x to $x + dx$ is $f_T(X)dX\ f_E(x)dx$. Assuming that $X_o = X + x$ we may write the probability

$$y_o = f_o(X_o)dX_o = \int_{-\infty}^{\infty} f_T(X_o - x)dX_o f_E(x)dx,$$

because $f_o(X_o)$ is obtained by taking into account that all possible values of x between $+\infty$ and $-\infty$ may be combined with a given X. This integral is of the same form as that given in Equation (2). Integration for the case where both $f_T(X)$ and $f_E(x)$ are normal gives

$$f_o(X_o) = \frac{1}{\sigma_o\sqrt{2\pi}} e^{-\frac{X_o^2}{2\sigma_o^2}}$$

where as before $\sigma_o = \sqrt{\sigma_T^2 + \sigma_E^2}$. This result is well known as the law of propagation of error.

When $f_E(x)$ is normal and $f_T(X)$ is given by the first two terms of the Gram-Charlier series, Equation (6), with skewness k_T and standand deviation σ_T, the observed distribution $f_o(X_o)$ is of the same functional form as the true distribution $f_T(X)$ and has values of standard deviation σ_o and skewness k_o given by Equations (5) and (8) in Part I. This result appears to be new.

Now for the case where the true distribution $f_T(X)$ and the law of error $f_E(x)$ are both second approximation type, the integration is somewhat tedious, but we can approach a special case of this problem easily from a slightly different angle as indicated in Appendix 2. Under certain special conditions therein set forth, the resultant distribution is also second approximation form with a skewness which is less than that of either $f_T(X)$ or $f_E(x)$ and is equal to $\frac{1}{\sqrt{2}} k_T$ when $k_T = k_E$, the standard deviation σ_o being again equal to $\sqrt{\sigma_T^2 + \sigma_E^2}$.

Example of Applications to Determine Most Economical Way of Measuring Quality

Let us next consider a very simple method of using the above results to indicate the most economical method for determining the quality of product with a given degree of precision.

What is the most economical way of determining the quality of product within some predetermined range $\overline{X} \pm \Delta\overline{X}$ with a known probability P, where $\overline{X}$ is the average quality? Let us assume that:

a_1 = cost of selecting each unit and making it available for measurement,

a_2 = cost of making each measurement,

n_1 = number of units selected,

n_2 = number of measurements made on each unit,

σ_1 = standard deviation of the errors of observation.

$\sigma_2 = \sigma_T$ = standard deviation of the true distribution $f_T(X)$.

Let us take $P = .9973$. Then the range $\overline{X} \pm 3\sigma_{\overline{X}}$ includes 99.73 per cent. of the observations, and hence $\Delta\overline{X} = 3\sigma_{\overline{X}}$.

The average of n_2 measurements made on one unit is the observed value of the magnitude X for that unit, and this average has the standard deviation $\sigma_E = \frac{\sigma_1}{\sqrt{n_2}}$. Hence, from the theory of the preceding section, the standard deviation of the observation is

$$\sigma_o = \sqrt{\sigma_T^2 + \sigma_E^2} = \sqrt{\sigma_2^2 + \frac{\sigma_1^2}{n_2}}.$$

The standard deviation of the average of n_1 observations is $\sigma_{\overline{X}} = \frac{\sigma_o}{\sqrt{n_1}}$ and we find upon solving for n_1,

$$n_1 = \frac{\sigma_2^2 + \frac{\sigma_1^2}{n_2}}{\sigma_{\overline{X}}^2}.$$

The cost of inspection is

$$y = a_1 n_1 + a_2 n_1 n_2,$$

and by customary methods this can be shown to be a minimum when

$$n_2 = \frac{\sigma_1}{\sigma_2}\sqrt{\frac{a_1}{a_2}}.$$

The following values correspond to one practical case:

$\Delta\overline{X} = .3$ unit	$a_1 = \$0.50$
$\sigma_1 = .3$ unit	$a_2 = \$0.02$
$\sigma_2 = .9$ unit	$P = .9973$

Thus with the aid of the above theory we find the most economical method of inspection requires 2 observations on each of 86 units.

Application in Setting Limit Lines

Over 99 per cent. of the averages of samples of size N drawn from a product whose law of distribution is $f_T(X)$ where $f_T(X)$ is either normal or second approximation may be expected to lie within the limits defined by the true average $\overline{X}$ plus or minus $3\frac{\sigma_T}{\sqrt{N}}$. If an average falls outside these limits, this fact is taken as probably indicating the existence of a trend or cyclic fluctuation in product, the cause of which should be sought. The presence of errors of measurement increases the separation of these limits to $6\sigma_o$ from $6\sigma_T$. Our precision of detecting trend or cyclic fluctuation is thereby decreased.

Cases often happen in practice where σ_o is from 15 per cent. to 25 per cent. greater than σ_T. In some instances σ_o has been found to be nearly 50 per cent. greater than σ_T.

PART III

Error Correction of Data Taken to Relate Observed Deviations in Quality of Product to Some Particular Cause

In many practical cases it is not possible to write down an equation to show how the quality of a finished product depends upon the factors controlled by different manufacturing steps. To cite one such case, we may know that the quality of the finished article depends upon the control of the temperature to which some of the piece parts are heated in the process of manufacture. Thus the microphonic properties of carbon depend upon the temperature to which the carbon is heated. In cases where the relationship between quality and some factor (such as temperature in the above illustration) can only be determined through a study of the correlation existing between the quality and the particular factor, use must be made of the correlation coefficient r which is defined as

$$r = \frac{\Sigma\, y\, x}{\sigma_x \sigma_y N}$$

where x and y represent respectively deviations from the average quality $\overline{X}$ and the average magnitude $\overline{Y}$ of some factor which is

to be controlled by the manufacturing process, and N is the number of observations. Now, if errors of observation are made in determining x and y, the observed correlation coefficient $r_{x_0y_0}$ is known to be given by the expression

$$r_{x_0y_0} = \frac{\sigma_x \sigma_y}{\sigma_{x_0} \sigma_{y_0}} r_{xy} \tag{10}$$

where $\sigma_{x_0} = \sqrt{\sigma_x^2 + \sigma_{x_E}^2}$ and $\sigma_{y_0} = \sqrt{\sigma_y^2 + \sigma_{y_E}^2}$,

σ_{x_E} and σ_{y_E} being the root mean square errors of observation of x and y respectively.

Attention is directed to Equation (10) which shows that the observed correlation coefficient $r_{x_0y_0}$ is always less than the true correlation coefficient r_{xy} irrespective of the number of observations made. Obviously, this point is of considerable commercial importance as we shall now see.

If the observed correlation is small, we customarily assume that there is little need of trying to control the quality X by controlling the manufacturing factor Y, whereas this conclusion cannot be justified unless it can be shown that the true correlation has not been masked by the errors of measurement.

This point has had to be taken into account in the development of machine methods for testing transmitters and receivers, because the calibration curves of the machines in terms of ear-voice tests depend upon the correlation coefficient.

Appendix I

It may be of some interest to certain readers to note that the results given in Equations (4) and (7) can also be obtained in the following way by the method of moments so often used in statistical investigations.

Assuming that $f_T(X+x)$ is expansible in terms of a Taylor's series, we get

$$f_o(X) = f_T(X) + \frac{\sigma_E^2}{2} f_T''(X) + \frac{1}{\underline{2}} \left(\frac{\sigma_E^2}{2}\right)^2 f_T''''(X) + \frac{1}{\underline{3}} \left(\frac{\sigma_E^2}{2}\right)^3 f_T^{vi}(X) + \ldots . \tag{11}$$

If we substitute a normal form for $f_T(X)$ in Equation (11) and solve for the moments of $f_o(X)$, we find that the odd moments are zero

and the ratio of the 4th moment to the square of the 2nd is numerically 3 which indicates that $f_o(X)$ is normal in form.

A similar substitution of the 2nd approximation form for $f_T(X)$ in Equation (11) yields a distribution $f_o(X)$ from whose moments we deduce Equation (7). Use is made in this proof of the easily demonstrated theorem that

$$\int_{-\infty}^{\infty} x^i f_E^j(x) = 0$$

if $i < j$, where f_E^j is the jth derivative of the normal law function.

APPENDIX II

It is well known that the normal law of distribution may result from a system of n (n being large) causes each of which produces an increment ΔX measured from some fixed origin with a probability $p = \frac{1}{2}$ and no increment with a probability $q = \frac{1}{2}$. Furthermore the second approximation may result from a similar system in which $p+q$ and n is large. Under such systems of causes, the probabilities of the occurrences of n, $n-1$, . . . 3, 2, 1, 0 increments are given by the successive terms of the point binomial $(p+q)^n$.

Let us assume that the symbols p_T, q_T, n_T, ΔX and p_E, q_E, n_E, Δx refer to the systems of causes controlling the product and errors respectively. The probabilities of observed combinations $n_T\Delta X + n_E\Delta x$, $(n_T-1)\Delta X+(n_E-1)\Delta x$, . . . are given by the successive terms of the expansion $(p_T+q_T)^{n_T}$ $(p_E+q_E)^{n_E}$. Now for the special case $p_T = p_E = p$ and $\Delta X = \Delta x$ we have the resultant probability distribution $(p+q)^{n_T+n_E}$ with skewness

$$k_o = \frac{q-p}{\sqrt{pq(n_T+n_E)}}$$

and standard deviation

$$\sigma_o = \sqrt{pq(n_T+n_E)}.$$

Now if $p = q$, the skewness k_o is zero and the observed distribution is more nearly normal than either component, and its standard deviation σ is the square root of the sum of the squares of σ_T and σ_E. This result is similar to that given by Equation (4) of this paper.

We may also consider by this method a case not treated in this paper. When the skewness k_T of the true values is equal to that k_E of the law of error, or, more particularly, when $n_T = n_E = n$, $p_T = p_E = p$, $q_T = q_E = q$, $p = q$, we see that the observed distribution is given

by the successive terms of $(p+q)^n$ and the skewness of the observed distribution k_o is $\frac{1}{\sqrt{2}}\ k$, and the standard deviation σ_o is $\sqrt{2}\ \sigma$; i.e. the observed skewness is only $\frac{1}{\sqrt{2}}$ times that of either the true distribution or the law of error, and the observed standard deviation σ_o is $\sqrt{2}$ times the standard deviation of either of the true or error distributions.

The Theory of the Operation of the Howling Telephone with Experimental Confirmation

By HARVEY FLETCHER

SYNOPSIS: A general theory of the sustained oscillations of electro-mechanical systems is presented in the paper. The electrodynamical properties of the telephone transmitter and receiver are described and sufficient numerical data are given to enable one to calculate the intensity and frequency of howling for various types of systems. Detailed consideration is given to the following three systems, namely, one where the transmitter and receiver disphragms are coupled together mechanically by a lever system, one where they are coupled by a small box of air, and one where they are coupled by a long tube of air. The type of electrical circuit to use with each of these systems depends upon the type of performance desired.

WHEN the telephone receiver of a subscriber's set is held in front of the mouthpiece of the transmitter, a shrill note is emitted. A sustained oscillation is set up in the electro-mechanical system which is frequently called "howling" or "singing" or "humming."

This phenomenon was first observed by A. S. Hibbard of the United States in 1890. Frank Gill was the first to publish an account of the phenomenon. He first noted that the pitch of the howling note was changed by reversing the telephone receiver connection. In summarizing further his experimental results, he states "that the pitch of the note appears to be determined by the length of the column of air between the two diaphragms and the conditions of the circuit. As the periodic time of the circuit is increased, the time of the note rises. To some extent, the pitch is governed by the rate of the diaphragm, but I do not think this is so important a factor as the others. The main factors appear to be the angle of lag and the length of the column of air between the diaphragms. Although the vibration is a forced one, we could almost see that its rate is largely dependent on the free period of the circuit." [1]

In 1908 Kennelly and Upson extended Gill's work and made extensive experimental investigations of the case in which the transmitter and receiver are coupled together acoustically by means of a

[1] Taken from a paper on "Notes on the Humming Telephone" by F. Gill, read at a meeting of the Dublin Local Section of the Society of Telephone Engineers and published in the Journal of the Institution of Electrical Engineers, Vol. XXXI, 1901.

hollow circular tube of varying lengths and electrically by means of an induction coil. The summary of the conclusions is as follows: [2]

"(1) The mean frequency of the humming-telephone note is determined solely by the receiver diaphragm, and its natural free rate of vibration. (2) The ascending intersections of the frequency zig-zag with the mean frequency line will be formed approximately at tube lengths of $(3/4+m)\ v/n_o$ cm. for one connection, and of $(1/4+m)\ v/n_o$ cm. for the other connection, of the receiver; where v is the velocity of sound in air, n is the mean frequency in cycles per second, and m is any positive integer, within the working range of the tube. The constants 3/4 and 1/4 may be modified by the presence of condensers, and other circumstances. (3) The range of pitch variation. and the breaking positions, are determined by the transmitter, and by the reinforcing capability of the system. For systems that are weak, either electrically or acoustically, the range of pitch, above or below the mean, will be small. (4) The primary current, as measured by a DC instrument, is ordinarily a minimum at the mean frequency, and a maximum at a break. (5) Transmitters may be tested for effectiveness, by measuring their hum-extinguishing resistances in the primary or secondary circuit. The tube length should be such as to produce mean frequency if one connection of receiver only is used, but should favor both connections equally, if both connections of receiver are used."

They also give a first approximation theory to account for the changes in frequency as the length of the coupling tube is changed.

In 1917, H. W. Nichols gave the general equations for the special case where the two diaphragms act as pistons closing the ends of a tube of air. This case was given as an illustrative example of the "Theory of Variable Dynamical Electrical Systems." [3]

This paper gives a theoretical treatment of the behavior of a system containing a transmitter and a receiver coupled together acoustically and electrically, and with a source of electrical energy feeding the transmitter. Formulae are deduced which give the frequency and intensity of howling in terms of the physical constants of the system. Numerical calculations are given and sufficiently detailed solution of some special cases are given to enable one, who is interested in using the howling telephone as a source of alternating current or for other experimental work, to design the set for his particular purpose.

[2] "Humming Telephone" by A. E. Kennelly and Walter L. Upson, American Philosophical Society, July 20, 1908.

[3] Physical Review, Aug., 1917, p. 191.

General Solution of the Howling Circuit

The elements of a telephone system which is howling are the transmitter, the receiver, the mechanical coupler and the electrical coupler as indicated in Fig. 1. If there is a source of electrical power in the electrical coupler, which is released by movements of the transmitter diaphragm in the form of electrical vibrations, and also, if there is a proper relationship between these four elements, then a sustained

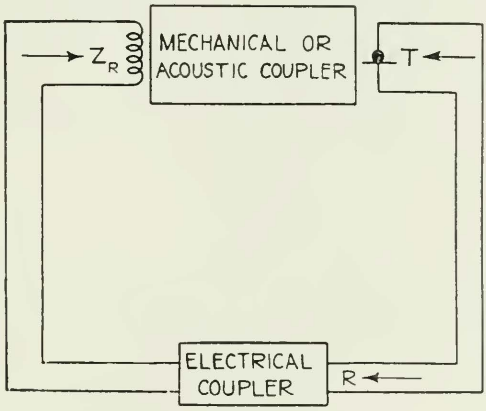


Fig. 1

howling will result. In other words, if the gain in the transmitter due to its amplifying action is just equal to the losses in the electrical and mechanical circuits, then a steady oscillatory state will be maintained. The problem is to determine the nature of these relationships.

Assume that the conditions are such that a steady oscillatory state has been set up. Under such conditions let T be the electrical impedance of the transmitter, R the impedance looking away from the transmitter terminals into the electrical coupler, and Z_R the impedance of the receiver. It is well known that the impedance Z_R is dependent upon the velocity of motion of the receiver diaphragm. Also, T is dependent upon the amplitude of motion of the transmitter diaphragm as well as upon the direct current supplied to it. Consequently, the impedances defined above are not only dependent upon frequency but also upon the mechanical coupling and magnitude of the current supplied to the transmitter.

If e is the electromotive force created in the transmitter, and i the

current flowing through it both expressed in root mean square values[4], then

$$e=(T+R)i \tag{1}$$

It is convenient to define a quantity M which I shall call the unilateral mutual impedance by the qeuation

$$e_1=Mi_1 \tag{2}$$

where e_1 is the electromotive force created in the transmitter when a current i_1 flows in the receiver circuit. It is a quantity which is closely related to the effectiveness of the mechanical coupling and the efficiencies of the transmitter and receiver.

If the electrical coupler be considered part of the receiver, and the transmitter and receiver circuits are connected together as in Fig. 1, then $e=e_1$, and $i=i_1$. Consequently

$$M=T+R \tag{3}$$

is the condition for sustained oscillation. This condition is in effect a pair of conditions, as the two sides of the equation must be equal both in amplitude and in phase. These two conditions are sufficient to determine the frequency and intensity of howling.

In order to express M and R in more fundamental physical constants, it is necessary to examine more closely the mechanical and electrical connections. Before doing this for some important special cases, it will be necessary to discuss some of the electro-dynamical properties of transmitters and receivers.

Electrodynamical Properties of Transmitters and Receivers

For the sake of clarity the discussion will be confined to permanent magnet receivers and carbon transmitters. The modifications necessary for other types of instruments will, I think, be evident from the discussion. Representing by F_R and F_T the forces acting on the diaphragms of the receiver and transmitter respectively, and by y and z their displacements, we have the following equations defining the "stiffness factors" S_R and S_T

$$S_R=\frac{F_R}{y} \tag{4}$$

$$S_T=\frac{F_T}{z} \tag{5}$$

[4] In what follows all quantities involving periodic variations will be expressed as root mean square values unless otherwise specified, and the vector notation will be used for denoting phases.

These factors are usually complicated functions of the frequency while S_T likewise depends on the kind and amount of agitation. In the case of a system of a single degree of freedom which may be regarded as a first approximation to this case

$$S = m\omega^2 + j\omega r + s \tag{6}$$

where ω is 2π times the frequency. When referring to the movements of a diaphragm, the quantity m represents the mass, r the mechanical resistance, and s the elastic constant. The stiffness factor S divided by $j\omega$ is usually called the mechanical impedance.

Measurements have shown that for the transmitters and the

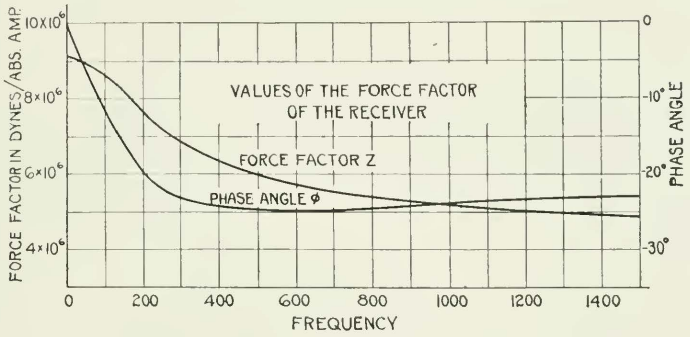


Fig. 2

receivers used in the experiments described below, the following constants represent approximately the two stiffness factors in the region of resonance

$$S_R = -.93\omega^2 + 230j\omega + 3\times 10^7 \tag{6'}$$

$$S_T = -4.5\omega^2 + 2000j\omega + 2\times 10^8 \tag{6''}$$

An important constant which enters into the determination of the unilateral mutual impedance M is the force factor of the receiver which will be designated by Z. It is defined as the force in dynes acting upon the diaphragm per unit of current. For the receivers used in this investigation, its values in magnitude and phase are shown for various frequencies in Fig. 2. These were determined by the method outlined by Wegel.[5] In the region of the resonant frequency its value in absolute units can be approximately represented by

$$Z = 5.3\times 10^6 \lfloor \overline{24^\circ}. \tag{7}$$

[5] Theory of Telephone Receivers—Wegel, R. L., Jour. of A. I. E. E., Oct. 1921.

The impedance Z_R of the receiver varies with frequency and depends upon the load on the diaphragm. If S is the loaded stiffness of the diaphragm, that is, its resistance to force under actual working conditions, and Z_d is the impedance of the receiver when the diaphragm is prevented from moving, then it is well-known that

$$Z_R = Z_d + j\frac{\omega Z^2}{S}. \tag{8}$$

It was found that Z_d expressed in ohms could be represented in the frequency region near resonance by the formula

$$Z_d = 93 + .06f + j(43 + .15f) \tag{9}$$

where f denotes the frequency in cycles per second.

The electromotive force e created in the transmitter, the direct current I flowing through it, and the displacement of the diaphragm are related in a rather complicated way. For describing this relationship it is convenient to define a modulation factor h by the equation

$$e = Ihz \tag{10}$$

Combining this equation with (2) it is seen that

$$M = Ih\frac{z}{i} \tag{11}$$

which shows that the modulation factor is also an important one in determining the unilateral mutual impedance. For a sustained oscillation the factor Ih does not enter into the periodic variation and may be thought of as an electro-mechanical impedance between the electromotive force created in the button and the displacement of the diaphragm of the transmitter. However, for a different condition of sustained oscillation which results in giving z a different magnitude the value of h changes. In other words h is dependent upon the agitation of the carbon as represented by z, and also upon the direct current supplied to the transmitter. It is mainly this variable character of h that makes it possible to fulfill the conditions for sustained howling.

Simultaneous measurements of e, I and z were made upon several transmitters of the type used in this investigation. From the results obtained and from the defining equation (10) for h, it was found that

the following empirical equation would represent approximately the relation between h, I and z, namely

$$h = \frac{32+\frac{z}{2}}{\left(2.6+2z+\frac{1}{z}\right)(I+.03)} \tag{12}$$

where z is expressed in microns and I in amperes e in volts and h in ohms per micron. To facilitate solving for z when h and I are given, a set of curves showing this relation is given in Fig. 3. It is this modulation factor h which measures the efficiency of the transmitter button.

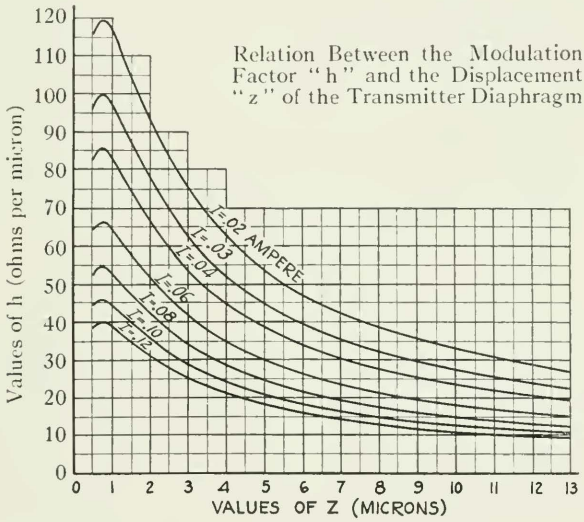


Fig. 3

It is also necessary to know the dependence of T upon z and I. To obtain this relation corresponding values of e and V, the DC drop across the transmitter as measured by direct current measuring instruments, were obtained for various degrees of agitation and amounts of direct current. Four transmitters were used in establishing the relation, the results being shown in Fig. 4. Then, for any value of the supply current I a value of T can be obtained from V. From the corresponding e a value of h and z can be obtained from equations (10) and (11). In this way the relations shown in Figs. 5

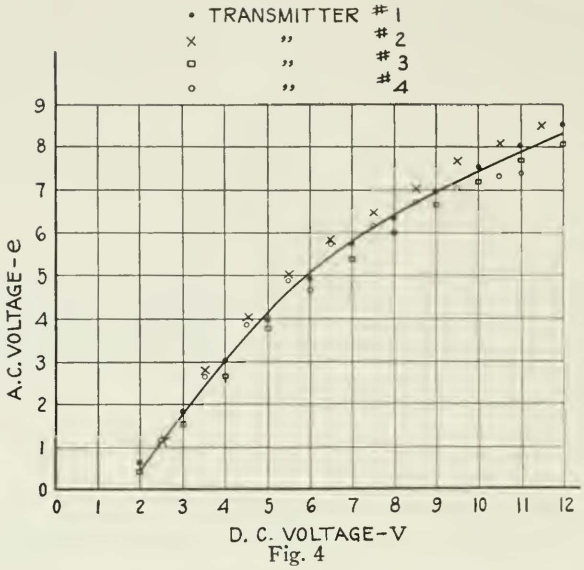

• TRANSMITTER #1
× " #2
□ " #3
○ " #4
A.C. VOLTAGE-e
D. C. VOLTAGE-V

Fig. 4

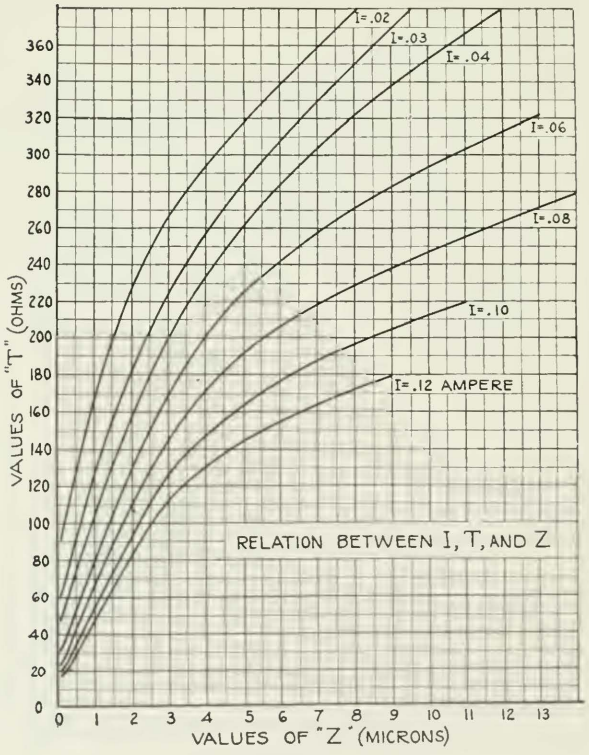

I=.02
I=.03
I=.04
I=.06
I=.08
I=.10
I=.12 AMPERE
RELATION BETWEEN I, T, AND Z
VALUES OF "T" (OHMS)
VALUES OF "Z" (MICRONS)

Fig. 5

and 6 were obtained. It is thus seen that for a given type of transmitter if the direct current and any one of the four quantities e, h, z, or T are known, the others are determined and may be obtained from suitable curves.

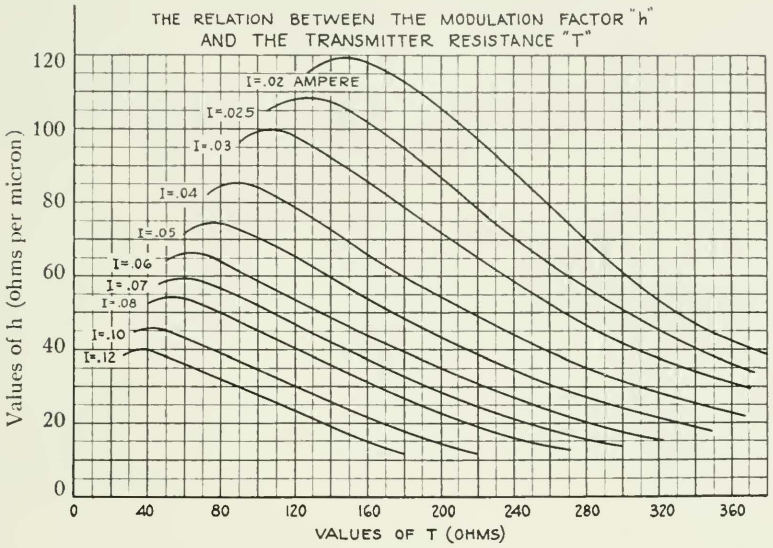


Fig. 6

Commercial receivers and transmitters have constants which vary largely from those given above. These values represent the general behavior of such instruments and are useful in understanding their operation in a howling circuit. Inasmuch as the performance of such instruments particularly the transmitter depends very largely upon the condition of operation the constants given cannot be applied with confidence to conditions greatly different from those mentioned in the paper. With these facts concerning telephone instruments in mind we are now in a position to treat some special cases.

Case 1—Diaphragms Connected Mechanically by a Rigid and Weightless Lever

To illustrate the method of solution this special case will be solved in some detail. A diagrammatic sketch illustrating the connections is shown in Fig. 7. Neglecting the reaction of the air, the vibration of the receiver diaphragm is controlled by the force Zi exerted by the

receiver winding and the opposing force X exerted by the connecting rod.

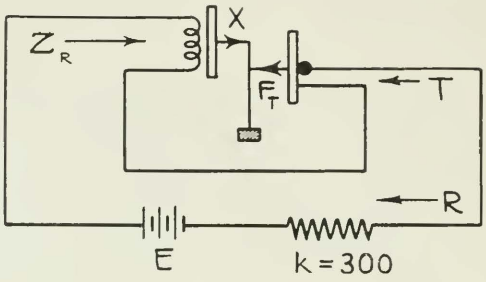


Fig. 7

The amplitude of motion of the receiver diaphragm is then given by

$$y = \frac{Zi - X}{S_R} \tag{13}$$

If the lever is rigid and weightless and has an arm ratio c, then

$$F_T = cF_R \tag{14}$$

and due to the restraint

$$y = cz = \frac{c^2 X}{S_T} \tag{15}$$

Using these equations together with equation (11) it is seen that

$$M = \frac{IhZ}{cS_R + \frac{1}{c} S_T} \tag{16}$$

$$S = S_R + \frac{1}{c^2} S_T, \tag{17}$$

$$R = Z_d + j\frac{\omega Z^2}{S} + k. \tag{18}$$

The relation between I and T is given by

$$I = \frac{E}{T + R_{DC} + k} \tag{19}$$

where R_{DC} is the direct current resistance of the receiver winding and k is the line resistance. The condition (3) for howling then becomes

$$IhZ = (Z_d + k + T)\left(cS_R + \frac{1}{c} S_T\right) + j\omega Z^2 c. \tag{20}$$

This is equivalent to two scalar equations and taken together with (19) and the curves of Fig. 6 gives the necessary four equations to solve for the unknowns f, h, T, and I.

The solution, however, is not straightforward since the relation between h, T, and I is only given empirically by a set of curves. By "cut and try" methods the solution for any numerical case can be obtained. The last term of (20) is usually negligible or at least it is of second order of magnitude. Consequently, the sum of the phase angles of the other factors must be approximately equal to the phase of Z. This completes the formal solution for this case.

The solution of a numerical case throws considerable light upon the physical phenomenon taking place, and also upon the method of calculation. Let the arm ratio be unity, a case corresponding to that when the diaphragms are connected directly together, and assume that the supply current is furnished by a battery of 24 volts through a line having a resistance of 300 ohms. Using the constants for the receivers and transmitters given above and expressing f in kilocycles, T in ohms, I in amperes and h in ohms per micron, equations (19) and (20) become

$$I = \frac{24}{384 + T'} \tag{19'}$$

$$Ih\ 52 \angle 24^\circ = [393 + T + 60f + j(43 + 150f)]\ [-2.14f^2 + 2.3 + j.14f] + j\ 1.7f. \tag{20'}$$

If I is positive there is no solution for f, since the angle of the first factor is in the first quadrant, and that of the second factor either in the first or second; consequently, the phases cannot match at any frequency. If the supply current is reversed, then I is negative or 180° is added to the phase of the left hand member making it a positive 156°. The solution for this case is

$f = 1072$ cycles	$i = 8.2$ mils
$h = 64$	$e = 5.5$ volts
$T = 150$ ohms	$y = z = 1.9$ microns
$I = 45$ mils	

If a value of c equal to 2.7, which is approximately equal to the square root of the ratio of mechanical impedances of the two diaphragms, then the solution for reversed DC supply becomes

$f = 1001$ cycles	$i = 10$ mils
$h = 47.3$	$e = 7.16$ volts
$T = 236$ ohms	$z = 3.9$ microns
$I = 39$ mils	$y = 10.5$ microns

It is thus seen that changing the ratio arm has increased the howling intensity, but the increase for the various elements is greatly different. The frequency is slightly lowered, the values of h and I have been

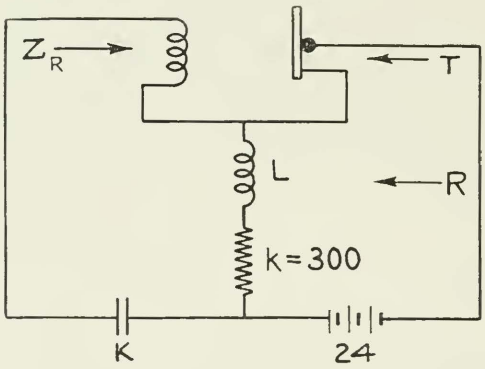


Fig. 8

reduced by 26% and 14% respectively, while the values of y, z, T, i and e have been increased 400%, 105%, 57%, 22% and 30% respectively.

If the circuit of Fig. 7 is modified as shown in Fig. 8, the inductance L being very large, then the condition for howling becomes

$$\frac{IhZ}{cS_R + \frac{1}{c}S_T} = T + Z_d + \frac{j}{K\omega} + j\frac{\omega Z^2}{S} \tag{21}$$

and

$$I = \frac{24}{300 + T}. \tag{22}$$

Using the same constants as above the condition for howling becomes

$$Ih\ 52\ |\overline{24^\circ}. = \left[93+T+60f+j(43+150f)-\frac{158}{Kf}\right][-2.14f^2+2.3+j.14f]+j1.7f \qquad (23)$$

The solution for values of $K=1$ *mf*, $K=1/2$ *mf*, and $K=1/5$ *mf* are given in Table I. When $K=1$ *mf* and the supply current is direct the solution which satisfies the phase equality is $f=506$. This corresponds to $h=220$ which is an impossible value. Therefore, no howling will be sustained for this condition. For $K=1/2$ *mf* the system will howl for both direct and reversed supply current, the frequency changing suddenly from 839 to 1119 cycles as the current is reversed while the other variables change only slightly.

TABLE I

	$K=1$		$K=1/2$		$K=1/5$	
	Direct	Reversed	Direct	Reversed	Direct	Reversed
f	.....	1016	839	1119	935	.....
h	220	33.5	53.5	57.4	44.2	.....
T	.....	275	160	140	220	.....
I	.....	42	52.2	54.5	46	.....
i	.....	20	18.3	17.6	10.9	.....
e	.....	8.2	6.15	5.6	7.5	.....
x	.....	5.9	2.2	1.8	3.7	.....
y	.....	16	5.95	4.9	10	.....

It is interesting to note the change in the howling frequency as the value of K increases. When the supply current is negative, and for values larger than 1 *mf*, the frequency of howling is always close to 1000, as K goes from 1 to 1/2 the frequency increases to above 1100. For smaller values of K the frequency continues to slowly increase until, for values smaller than 1/3, the system ceases to sustain oscillations. For positive values of supply current no howling will result until K becomes smaller than 2/3 where the frequency is around 800. The frequency then increases reaching a howling frequency around 1000 for $K=1/7$. For smaller values of K no howling will be sustained.

Case II—Diaphragms Coupled Together by a Small Chamber of Air

It will be assumed that the air chamber is so small that the phase of the pressure variation is the same on both diaphragms. Let V be the volume of air between the diaphragms. Then

$$V = V_0 + Q_R y + Q_T z \tag{24}$$

where V is the volume of air in the undisturbed state and Q_R and Q_T are the effective areas of the receiver and transmitter diaphragms respectively.

The pressure variation in the chamber (changes considered adiabatic) is given by

$$dp = -\gamma \frac{P}{V} dV = -(Q_R y + Q_T z)\gamma \frac{P}{V} \tag{25}$$

When the steady state is set up this may be considered a vector equation and the variables expressed in *rms* values.

The equations of motion for the diaphragms are

$$y = \frac{Zi - Q_R dp}{S_R} \tag{26}$$

and

$$z = \frac{Q_T dp}{S_T}. \tag{27}$$

Solving

$$y = \frac{Zi\left(S_T + \frac{\gamma P}{V} Q_T^2\right)}{S_R S_T + \frac{\gamma P}{V} Q_T^2 S_R + \frac{\gamma P}{V} Q_R^2 S_T}, \tag{28}$$

$$z = -\frac{\frac{\gamma P}{V} Q_R Q_T}{S_T + \frac{\gamma P}{V} Q_T^2} y, \tag{29}$$

$$M = \frac{IhZQ_R Q_T}{\frac{V}{\gamma P} S_R S_T + Q_T^2 S_R + Q_R^2 S_T}. \tag{30}$$

In this case the ratio between z and y is not fixed, but depends upon S_T which is a function of the frequency.

The loaded stiffness of the receiver diaphragm is

$$S=\frac{\frac{V}{\gamma P}S_R S_T+Q_T{}^2S_R+Q_R{}^2S_T}{\frac{V}{\gamma P}S_T+Q_T{}^2}. \tag{31}$$

For the transmitter and receiver used

$$Q_R=6.5,$$
$$Q_T=10.3.$$

Let the volume of entrapped air be taken as 10 cc., then

$$\frac{\gamma P}{V}=1.418\times 10^5.$$

Using these values and the values for S_R and S_T and the circuit of Fig. 8 with $K=\frac{1}{2}$ the condition for howling becomes

$$\begin{aligned} Ih\,3.48\,\underline{|24^\circ} &= 27.6f^5+(50.3+.459T)f^4-59.9f^3-(1.01T+85.7)f^2+31.9 \\ &\quad +(.539T+33)+2\left[68f^5+16.7f^4-(.0506T+11.1)f^3-40f^2\right. \\ &\quad \left.+(.0537T-233)f+23.2+\frac{172}{f}\right] \end{aligned} \tag{32}$$

where I is expressed in amperes, T in ohms, f in kilocycles and h in ohms per micron.

For reverse current or negative I the solution is

$f=970$ kilocycles	$i=24\,\underline{|17^\circ}$
$h=30.5$	$e=8.7$ volts
$T=290$ ohms	$z=7.0$ microns
$I=.0407$ mils	$y=1.9\,\underline{|158^\circ}$ microns

Comparing this to the case where the diaphragms are coupled by a lever having an arm ratio 2.7 it is seen that the air coupling produces a greater e.m.f. in the transmitter and only a slightly increased AC current. The receiver diaphragm in this case, however, has a smaller amplitude than the transmitter diaphragm. At this particular howling frequency the transmitter diaphragm stiffness is only about 1/4 that of the receiver diaphragm stiffness which explains this anomalous result. Also, it will be seen that the diaphragms vibrate almost oppositely in phase.

These cases are sufficient to illustrate the method of calculation, but there is one other important case for which I desire to give the results as this is the case handled experimentally by Kennelly and Upson.

Case III—Diaphragms Connected Acoustically by a Tube of Air of Uniform Cross-Section with an Air Chamber at Both Ends

In this case the two diaphragms are connected acoustically by the air, but since the tube has considerable length phase differences exist

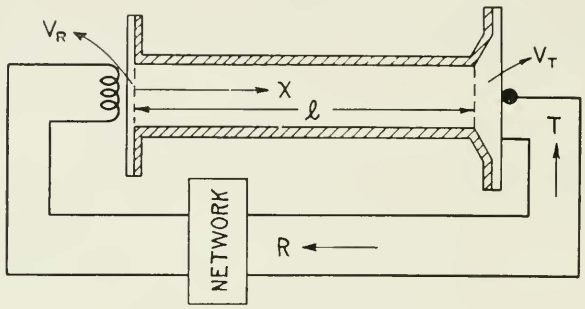


Fig. 9

at different points along it. The connections are shown schematically in Fig. 9.

The equation of motion for the receiver diaphragm is

$$y = \frac{Zi - Q_R dp_R}{S_R} = \frac{Zi}{S} \tag{33}$$

and for the transmitter diaphragm is

$$z = \frac{Q_T dp_T}{S_T} \tag{34}$$

where dp_R and dp_T are the pressure variations in the air chambers at the receiver and transmitter ends of the tube respectively.

The equations of motion for a gas in which the movements are small and in only one direction and in which the fluid friction is neglected are as follows:[6]

$$\frac{d^2\phi}{dt^2} = a^2 \frac{d^2\phi}{dx^2}, \tag{35}$$

$$\frac{dp}{\rho} = -\frac{d\phi}{dt}, \tag{36}$$

[6] See Rayleigh "Theory of Sound," Vol. II, pp. 14 and 15, 49 and 50.

where ϕ is the velocity potential, t the time, a the velocity of sound in the air, x the distance along the tube, p the pressure and ρ the density of the air.

For the case in which we are interested, a sinusoidal oscillation is sustained, so that the special solution

$$\phi = e^{j\omega t}\left(A\cos\frac{\omega X}{a} + B\sin\frac{\omega x}{a}\right) \tag{37}$$

is suitable for our problem. Quantities A and B are arbitrary constants which are determined by the end conditions. Substituting this value of ϕ in equation (35), there results

$$dp = -\rho j\omega e^{j\omega t}\left(A\cos\frac{\omega x}{a} + B\sin\frac{\omega x}{a}\right) \tag{38}$$

It remains then to determine the arbitrary constants A and B.

At the receiver end of the tube, the displacement, ζ_R of the air diaphragm across the end of the tube is related to the displacement y of the receiver diaphragm. This relationship is established by the following consideration. If q is the cross-section of the tube, the increase in volume in the air chamber is given by

$$dV_R = (\zeta_R q - yQ_R). \tag{39}$$

Assuming that the air chamber is so small that the pressure change at any instant is the same throughout, and that it takes place adiabatically, we have:

$$dp_R = -\gamma\frac{p}{V_R}dV_R \tag{40}$$

Combining equations (33), (39), and (40), we obtain:

$$qS_R\zeta_R = Q_R Zi - \left(\frac{V_R}{\gamma p}S_R + Q_R{}^2\right)dp_R \tag{41}$$

Similarly,

$$qS_T\zeta_T = \left(\frac{V_T}{\gamma p}S_T + Q_T{}^2\right)dp_T \tag{42}$$

Then the following conditions must be fulfilled at the two ends of a tube of length l.

At $x=0$, $\qquad dp = dp_R$ and $\dfrac{d\phi}{dx} = \dfrac{d\zeta_R}{dt}$;

at $x=l$, $\qquad dp = dp_T$ and $\dfrac{d\phi}{dx} = \dfrac{d\zeta_T}{dt}$.

These conditions give the following equations:

$$a\omega\rho A + S_R' B = jaZ' i_O \tag{43}$$

$$\left(a\omega\rho \cos\frac{\omega l}{a} + S_T' \sin\frac{\omega l}{a} \right) A + \left(a\omega\rho \sin\frac{\omega l}{a} - S_T' \cos\frac{\omega l}{a} \right) B = 0 \tag{44}$$

where

$$S_R' = \frac{qS_R}{Q_R^2 + \frac{V_R}{\gamma p} S_R}, \quad S_T' = \frac{qS_T}{Q_T^2 + \frac{V_T}{\gamma p} S_T}, \quad Z' = \frac{ZQ_R}{Q_R^2 + \frac{V_T}{\gamma p} S_R}.$$

Solving for the constants A and B, we find their values to be:

$$A = jaZ' i_o \left(S_T' \cos\frac{\omega l}{a} - a\omega\rho \sin\frac{\omega l}{a} \right) \div D, \tag{45}$$

$$B = jaZ' i_o \left(S_T' \sin\frac{\omega l}{a} + a\omega\rho \cos\frac{\omega l}{a} \right) \div D, \tag{46}$$

where

$$D = [S_R' S_T' - (a\omega\rho)^2] \sin\frac{\omega l}{a} + a\omega\rho (S_R' + S_T') \cos\frac{\omega l}{a}. \tag{47}$$

The two pressure values are then given by:

$$dp_R = Z_1' a\omega\rho \left(S_T \cos\frac{\omega l}{a} - a\omega\rho \sin\frac{\omega l}{a} \right) \div D, \tag{48}$$

$$dp_T = Z_1' a\omega\rho \, S_T' \div D, \tag{49}$$

and

$$y = \frac{Zi}{S_R D} \left[S_R' S_T' - (a\omega\rho)^2 \left(1 - Q_R \frac{Z'}{Z} \right) \sin\frac{\omega l}{a} + a\omega\rho \left(S_R' S_T' \left(1 - Q_R \frac{Z'}{Z} \right) \right) \cos\frac{\omega l}{a} \right], \tag{33'}$$

$$z = \frac{qZia\omega\rho}{D} \left[\frac{Q_T}{Q_T^2 + \frac{V_T}{\gamma p} S_T} \; \frac{Q_R}{Q_R^2 + \frac{V_R}{\gamma p} S_R} \right]. \tag{34'}$$

The loaded stiffness of the receiver diaphragm is given by

$$S = \frac{qQ_T Q_R a\omega\rho \left(N \sin\frac{\omega l}{a} + P \cos\frac{\omega l}{a} \right)}{S_T \frac{a\omega\rho}{\gamma p} \left[\left((q^2 - \frac{a\omega\rho}{\gamma p} V_R V_T) \sin\frac{\omega l}{a} + q(V_T + V_R) \cos\frac{\omega l}{a} - a\omega\rho \left(\frac{a\omega\rho}{\gamma p} V_R Q_R^2 \sin\frac{\omega l}{a} + qQ_T \cos\frac{\omega l}{a} \right) \right.} \right., \tag{50}$$

where

$$N = S_T S_R \frac{1}{a\omega\rho}\left[\frac{q^2}{Q_R Q_T} - \frac{(a\omega\rho)^2}{(\gamma p)^2}\frac{V_T V_R}{Q_R Q_T}\right] - S_R\left(\frac{a\omega\rho}{\gamma p}\right)\frac{Q_T}{Q_R} V_R - S_T\left(\frac{a\omega\rho}{\gamma p}\frac{Q_R}{Q_T} V_T, \quad (51)$$

$$P = S_R \frac{Q_T}{Q_R} + S_T \frac{Q_R}{Q_T} + S_T S_R \frac{V_T + V_R}{Q_T Q_R}\frac{1}{\gamma p}. \quad (52)$$

The unilateral mutual impedance M is given by

$$M = \frac{IhZ}{N \sin\frac{\omega l}{a} + P \cos\frac{\omega l}{a}}. \quad (53)$$

The condition for sustained howling becomes

$$\frac{IZ}{T+R}h = N \sin\frac{\omega l}{a} + P \cos\frac{\omega l}{a}. \quad (54)$$

If the two diaphragms work directly into the connecting tube as pistons, then $Q_R = Q_T = q = Q$ and $V_R = V_T = O$ and the expressions for M and S become [7]

$$M = \frac{IhZ\, Q\, a\omega\rho}{[S_R S_T - (a\omega\rho)^2 Q^2] \sin\frac{\omega l}{a} + (S_R + S_T) Q\, a\omega\rho \cos\frac{\omega l}{a}}, \quad (55)$$

$$S = \frac{[S_R S_T - (a\omega Q)^2] \sin\frac{\omega l}{a} + (a\omega\rho Q)\,(S_R + S_T)\cos\frac{\omega l}{a}}{S_T \sin\frac{\omega l}{a} + a\omega\rho\, Q \cos\frac{\omega l}{a}}. \quad (56)$$

The method of solution is the same as that given for the simpler cases, although it is evident that the actual work of calculation is more involved.

It is seen that in such a system the intensity and frequency depend upon a large number of quantities, namely: S_T and S_R, the diaphragm stiffness factors; Q_R and Q_T the effective areas of the two diaphragms; V_R and V_T the volumes of air entrapped between the diaphragm and the opening into connection tube; the length l and the cross section q of the connecting tube; the pressure a, the density s, and the velocity of sound a for the gas in the connecting tube; the resistance T, direct current I and modulation factor h of the transmitter; and

[7] These two equations were given by H. W. Nichols in essentially this form in the Physical Review, Vol. 10, p. 171; 1917.

the force factor and impedance of the receiving circuit. Modification of any of these may produce marked changes in the resulting howling.

The way the length l enters the formula (54) for sustained howling indicates that the curves representing the possible frequencies of howling, that is, frequencies which produce equality of phase on both sides of the equation, vary periodically with the length.

The intersection of the branches of these curves on any given frequency line will be separated by distances corresponding to $\frac{a}{f}$, that is, corresponding to a wave length at the pitch corresponding to f. Also, if the supply current is reversed, that is, the sign of I

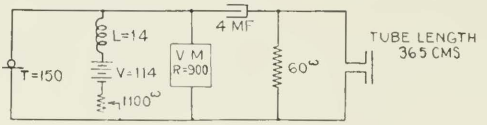


Fig. 10

changed, and the length of the tube varied until the frequency of howling is brought back to the original value, the change in length must be equal to $\frac{a}{2f}$. For since the frequency is unchanged all the quantities in equation (54) remain unchanged except the sine and cosine factors. Adding a half wave length is equivalent to adding π to the angle which makes the left hand member the negative of its first value, and consequently, restores the phase equality.

Using the circuit shown in Fig. 10 for the electrical coupling, the frequency of howling was computed for various tube lengths, the results being given in Fig. 11.

The instrument constants were those used before, the other values being $V_R = 1.6$ cc., $V_T = 6.4$ cc., and $q = .97$ cm.^{2}, $a = 3.43 \times 10^4$ cm/sec. $\rho = .001203$ gm/cm^3. Using these values the formulae for N and P become

$$N = (-1.31f^5 + 7.5f^3 - 9.68f + 3.26\frac{1}{f}) \times 10^8 + j(.141f^4 - .63f^2 + .36) \times 10^8,$$

$$P = (5.5f^4 - 12.35f^2 + 6.77) \times 10^8 + j(-.60f^3 + .66f) \times 10^8,$$

where f is the frequency in kilocycles.

The points on the calculated curves of Fig. 11 were obtained by direct experimental observation with the circuit shown, and with various lengths of brass tube coupling the transmitter and receiver together. The agreement between the calculated and observed

values is well within the experimental error involved in determining the constants used in the calculation.

In Fig. 12 are shown similar calculated curves for a transmitter called "hollow," that is, for one having a lower natural period of vibration. It is coupled to the same receiver as used before. The dotted curves in each case represent the behavior for reversed current.

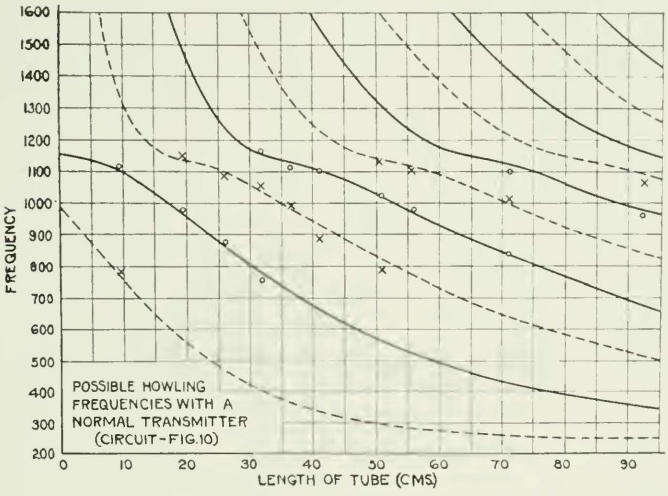


Fig. 11

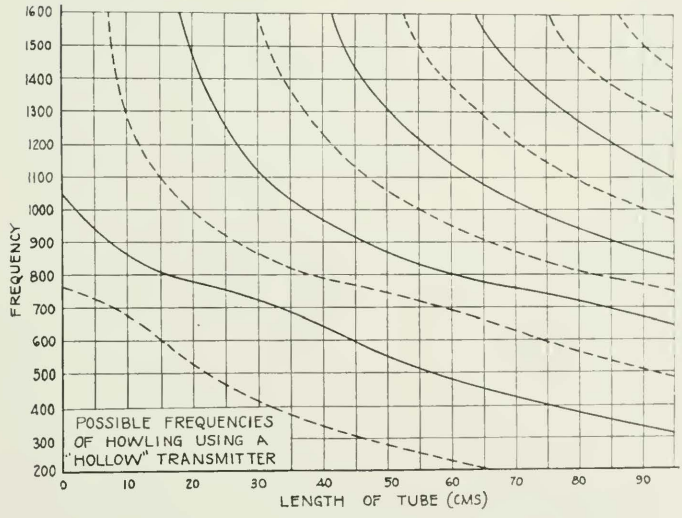


Fig. 12

In Figs. 13 and 14 are shown the probable frequencies of howling for these two transmitters as the tube length of the coupler is increased. The shaded areas are the so-called breaking points where the howling may be at either of the frequencies shown.

With these facts in mind let us review the conclusions reached by Kennelly and Upson given in the beginning of this paper. It is seen that conclusion (1) is not warranted. The transmitter and circuit conditions as well as the receiver diaphragm influence the mean frequency of humming. The second conclusion regarding the branches

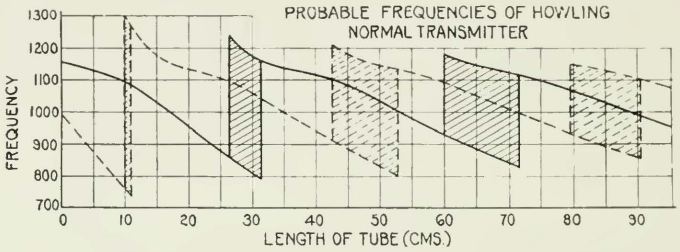


Fig. 13

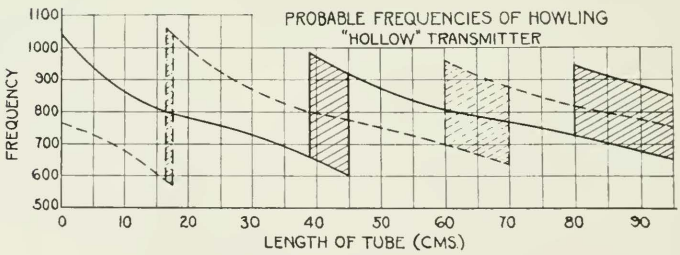


Fig. 14

of the curves representing the relation between frequency and tube length is correct and the explanation has just been given. This periodic relation is not only true of the mean frequency line but for every constant frequency line.

The terms corresponding to $\frac{1}{4}\frac{V}{n_o}$ and $\frac{3}{4}\frac{V}{n_o}$ depend upon a number of factors including the circuit and end conditions. Conclusion (3) is partially correct, the range of the howling frequencies depending upon the efficiencies of the transmitter, receiver, and circuit is evident from equation (54). Calculations show that conclusion (4) is generally correct although not necessarily so.

When the transmitter and receiver are coupled by the air in an open room the behavior is somewhat similar to the case just solved. The size and shape of the room as well as the disposition of articles of furniture will all influence the intensity and frequency of howling. In general when the two instruments are moved apart the frequency will go up and down similar to that when they are coupled by a tube.

NOMENCLATURE

T	Transmitter Resistance.
R	Impedance looking away from Transmitter Terminals.
Z_R	Impedance of Receiver.
Z_d	Damped Impedance of Receiver.
e	Electromotive Force Created in the Transmitter.
i	Alternating Current in the Transmitter Branch.
M	Unilateral Mutual Impedance between Receiver Current and Transmitter e.m.f.
F_R	Force on Receiver Diaphragm.
F_T	Force on Transmitter Diaphragm.
S_R	Stiffness Factor of Receiver Diaphragm.
S_T	Stiffness Factor of Transmitter Diaphragm.
y	Receiver Diaphragm Displacement.
z	Transmitter Diaphragm Displacement.
m	Mass of Diaphragm.
r	Mechanical Resistance of Diaphragm.
s	Elastic Constant of Diaphragm.
f	Frequency.
ω	2π times Frequency.
Z	Force Factor of Receiver.
S	Loaded Stiffness of Receiver Diaphragm.
I	Direct Current Supplied to Transmitter.
V	DC Voltage Drop across Transmitter Terminals.
h	Modulation Factor of the Transmitter.
X	Mechanical Force on Receiver Diaphragm for Case I.
E	Electromotive Force of Supply Battery.
k	Resistance in Line for Case I.
K	Capacity of Condenser.
V_R	Volume of Air in Front of Receiver Diaphragm.
V_T	Volume of Air in Front of Transmitter Diaphragm.
Q_R	Effective Area of Receiver Diaphragm.
Q_T	Effective Area of Transmitter Diaphragm.
p	Air Pressure.
γ	Adiabatic Constant.
ϕ	Velocity of Potential.
a	Velocity of Sound in Air.
x	Distance Along Connecting Tube.
ρ	Density of Air.
ζ_R	Displacement of Air Particle at Receiver End of Tube.
ζ_T	Displacement of Air Particle at Transmitter End of Tube.

Electric Circuit Theory and the Operational Calculus[1]

By JOHN R. CARSON

CHAPTER VI

Propagation of Current and Voltage Along the Non-Inductive Cable

THE principal practical applications of the operational calculus in electrotechnics are to the theory of the propagation of current and voltage along transmission systems. Of such transmission systems the simplest is the non-inductive cable. The theory of the non-inductive cable is not only of great historic interest, relating as it does to Kelvin's early work on the possibility of transatlantic telegraphy, but is also of very considerable practical importance today, and serves as a basis for the theory of submarine telegraphy over long distances. We shall therefore consider the propagation phenomena in the non-inductive cable in some detail.

The propagation phenomena in any type of transmission system are isolated and exhibited in the clearest possible manner when we confine attention to the infinitely long line, with voltage applied directly to the line terminals. Furthermore, as we shall see later, the solution for the infinitely long line is fundamental and can be extended to the more practical case of the finite line with terminal impedances. We therefore, in this chapter, shall confine our attention to the case of the infinitely long cable with voltage applied directly to the cable terminals.

Consider a cable of distributed resistance R and capacity C per unit length, extending from $x=0$ along the positive x axis. From a previous chapter (see equations (64) and (65)), we are in possession of the operational equations of voltage and current; they are, for the infinitely long line,

$$V=e^{-\sqrt{\alpha p}}\ V_o, \tag{162}$$

$$I=\frac{1}{Rx}\sqrt{\alpha p}\ e^{-\sqrt{\alpha p}}\ V_o=\sqrt{\frac{Cp}{R}}e^{-\sqrt{\alpha p}}\ V_o, \tag{163}$$

where $\alpha=x^2RC$, and V_o is the terminal cable voltage at $x=0$. Let us now assume that the terminal voltage V_o is a "unit e.m.f."; then

$$V=e^{-\sqrt{\alpha p}}, \tag{164}$$

$$I=\frac{1}{Rx}\sqrt{\alpha p}\ e^{-\sqrt{\alpha p}}. \tag{165}$$

[1] Continued from the October, 1925, issue.

The solution of (164) for V was considered in some detail in the preceding chapter; it is, by (129)

$$V=\frac{1}{\sqrt{\pi}}\int_0^{\tau}\frac{e^{-1/\tau}}{\tau\sqrt{\tau}}\,d\tau \tag{166}$$

where $\tau=4t/\alpha=4t/x^2RC$. Series expansions of this solution were also given. Another equivalent form is, by (131)

$$V=1-\frac{2}{\sqrt{\pi}}\int_0^{1/\sqrt{\tau}}e^{-\tau^2}\,d\tau. \tag{167}$$

This last form, recognizable also from inspection of the series expansion (132), is useful because the integral term is what is called the error function and has been completely computed and tabulated.

Before discussing these formulas and the light they throw on propagation phenomena in the non-inductive cable, we shall derive the solution for the current. A very simple way of doing this is to make use of the differential equation (57)

$$I=-\frac{1}{R}\frac{\partial}{\partial x}V.$$

Now from (166) and the relation

$$\frac{\partial}{\partial x}=\frac{d\tau}{dx}\frac{d}{d\tau}$$

we get

$$\frac{\partial}{\partial x}V=\frac{1}{\sqrt{\pi}}\frac{e^{-1/\tau}}{\tau\sqrt{\tau}}\frac{d}{dx}\frac{4t}{x^2RC}$$

$$=-\frac{2}{x\sqrt{\pi}}\frac{e^{-1/\tau}}{\sqrt{\tau}},$$

whence

$$I=\frac{2}{xR\sqrt{\pi}}\frac{e^{-1/\tau}}{\sqrt{\tau}}=\sqrt{\frac{C}{\pi Rt}}e^{-1/\tau}. \tag{168}$$

It is worthwhile verifying the formula by direct solution from the operational equation (165). From formula (g) of the table of integrals, we have

$$h=e^{-2\sqrt{\lambda p}}\sqrt{p}\,\sqrt{\frac{C}{R}}$$

$$=\frac{e^{-\lambda/t}}{\sqrt{\pi t}}\sqrt{\frac{C}{R}}.$$

Comparison with the operational equation shows that they are identical, within a constant factor provided we put $\lambda = \alpha/4$. Consequently the solution of (165) is

$$I = \sqrt{\frac{C}{\pi Rt}}\, e^{-\alpha/4t} = \sqrt{\frac{C}{\pi Rt}}\, e^{-1/\tau}$$

which agrees with (168). This, it may be remarked, is an excellent example of the utility of the table of integrals in solving operational equations.

This formula is easily calculated for large values of t by expanding the exponential function; it is

$$\frac{2}{Rx}\,\frac{1}{\sqrt{\pi\tau}}\left[1-\left(\frac{1}{\tau}\right)+\frac{1}{2!}\left(\frac{1}{\tau}\right)^2-\cdots\right].$$

The propagation phenomena of the non-inductive cable are therefore determined by the pair of equations

$$V = \frac{1}{\sqrt{\pi}}\int_0^{\tau}\frac{e^{-1/\tau}}{\tau\sqrt{\tau}}\,d\tau = 1-\frac{2}{\sqrt{\pi}}\int_0^{1/\sqrt{\tau}} e^{-\tau^2}\,d\tau \tag{169}$$

and

$$I = \frac{2}{\sqrt{\pi}xR}\,\frac{e^{-1/\tau}}{\sqrt{\tau}} = \sqrt{\frac{C}{\pi Rt}}\, e^{-1/\tau} \tag{170}$$

where $\tau = 4t/\alpha = \dfrac{4t}{x^2RC}$.

Now an important feature of these formulas is that the voltage at point x is a function only of $\dfrac{4}{x^2RC}\,t$; that is, of *4t divided by the total resistance and capacity* of the cable from $x=0$ to $x=x$. The same statement holds for the *form* of the current wave: its magnitude, however, is inversely proportional to xR, or the total resistance of the cable up to point x. Consequently a single curve, with proper time scale serves to give the voltage wave at any point on the cable. Similarly a single curve, with proper time and amplitude scales, serves to depict the current wave at any distance from the cable terminals. These curves are given in Figs. 3 and 4.

Referring to the curve depicting the current wave, we observe that it is finite for all values of $t>0$; consequently, in the ideal cable, the velocity of propagation is infinite. This is a consequence, of course, of the fact that the distributed inductance of the cable is neglected. Actually, of course, the velocity of propagation cannot exceed the

velocity of light. The error, however, in neglecting the inductance in the case of long cables is appreciable only near the head of the wave provided we confine attention to d.c. or low frequency voltages. This point will be discussed and explained more fully in connection with the transmission line.

The current, while finite, is negligibly small until τ reaches the

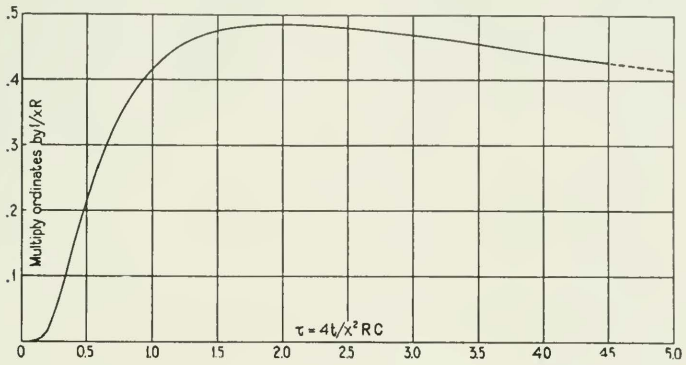


Fig. 3—Current in non-inductive cable ($G=0$) unit e.m.f. applied

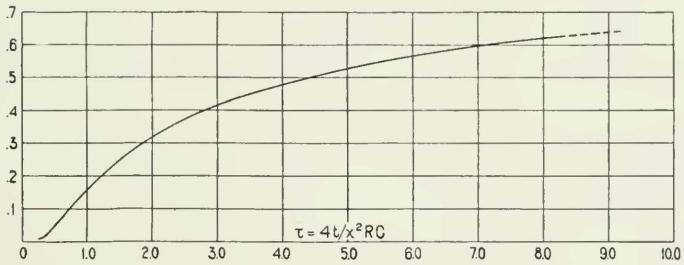


Fig. 4—Voltage in non-inductive cable ($G=0$) unit e.m.f. applied

value 0.2. In the neighborhood of this point it begins to build up rapidly; reaches at $\tau=2$ its maximum value

$$\frac{2}{\sqrt{\pi}\,xR}\,\frac{e^{-0.5}}{\sqrt{2}}=\frac{2}{\sqrt{\pi}\,xR}(0.429)$$

and then begins to decrease, ultimately dying away in accordance with the formula

$$\frac{2}{\sqrt{\pi}\,xR}\,\frac{1}{\sqrt{\tau}}\left\{1-\frac{1}{\tau}+\frac{1}{2!}\left(\frac{1}{\tau}\right)^2-\cdots\right\}.$$

Its subsidence to its final zero value is very slow; for example, when $\tau = 100$ its value is still

$$\frac{2}{\sqrt{\pi}\, xR}(0.10).$$

Turning to the voltage curve, Fig. 4, we see that it is negligibly small until τ reaches the value 0.25, at which point it begins to build up. Its maximum rate of building up occurs when $\tau = 2/3$, after

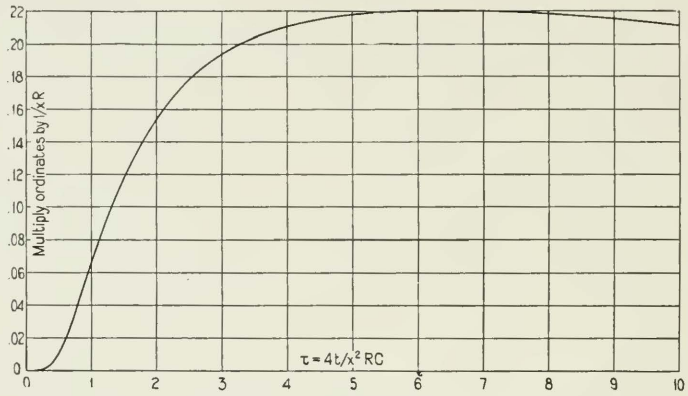


Fig. 5—Power transmitted in non-inductive cable ($G = 0$)

which it builds up more and more slowly. Its approach to its final steady value is in accordance with the formula

$$V = 1 - \frac{2}{\sqrt{\pi\tau}}\left(1 - \frac{1}{3\tau} + \frac{1}{2!}\frac{1}{5\tau^2} - \cdots\right).$$

Even, therefore, when τ is as great as 100, V differs sensibly from its ultimate value, unity, its value being 0.8876.

Since the actual time is $\frac{x^2RC}{4}\tau$, it follows that the speed of building up is inversely proportional to the square of the length of the cable.

The power curve VI is given in Fig. 5. V.I is the rate at which energy is being transmitted past the point x of the cable.

The fact that the form of the current and voltage waves depends only on $4t/x^2RC$ is at the basis of Kelvin's famous "KR" law, long applied to cable telegraphy and sometimes incorrectly applied to telephony. When the first transatlantic telegraph cable was under consideration, Kelvin attacked the problem of propagation along the non-inductive cable and arrived at formulas equivalent to (169) and

(170). From these formulas he announced the law that the "speed" of the cable, i.e., the number of signals transmissible per unit time, is inversely proportional to the product of the total capacity and total resistance of the cable (KR in the English notation). To see just what this means requires a little digression into the elementary theory of telegraph transmission.

Telegraph signals are transmitted in code by means of "dots" and "dashes." The "dot" is the signal which results when a battery is impressed on the cable for a definite interval of time, after which the cable is short circuited. A "dash" is the same except that the time interval during which the battery is connected to the cable is increased. The "dots" and "dashes" are separated by intervals, called "spaces", during which the cable is short circuited. Now when the cable is short-circuited we may imagine a negative battery impressed on the cable in series with the original battery. Consequently the current in the cable, corresponding to a signal composed of a series of dots, dashes and spaces, will be represented by a series of the form

$$I(t)-I(t-t_1)+I(t-t_2)-I(t-t_3)+I(t-t_4)-\ \ldots \qquad (171)$$

where, in the cable under consideration, $I(t)$ is given by (168). t_1 is the duration of the first impulse, t_2-t_1 of the first space, t_3-t_2 of the second impulse, etc.

Now by (168)

$$I(t)=\frac{2}{xR\sqrt{\pi}}\frac{e^{-1/\tau}}{\sqrt{\tau}}=\frac{2}{xR\sqrt{\pi}}\phi(\tau).$$

τ is, of course, $4t/x^2CR=4t/KR$ (in the English notation). Now suppose that

$$\tau_1=\frac{4t_1}{x^2CR},$$

$$\tau_2=\frac{4t_2}{x^2CR},\ \text{etc.}$$

Then the signal can be written as

$$\frac{2}{xR\sqrt{\pi}}\{\phi(\tau)-\phi(\tau-\tau_1)+\phi(\tau-\tau_2)-\ \ldots\ \} \qquad (172)$$

Now if the relative time intervals τ_1, τ_2 . . . are kept constant (as the length of the cable is varied), the actual time intervals t_1, t_2 . . . are proportional to x^2CR or to KR, and the wave form of the total signal is independent of KR, when referred to the relative time scale τ.

Hence, if T is the total time of the signal, T is proportional to x^2CR (or to KR). That is to say, if the duration of the component dots, dashes, spaces of the signal are proportional to the "KR" of the cable, the wave form of the received signal, referred to the τ time scale, is invariable, and the total time required to transmit the signal is proportional to the "KR" of the cable. Now the maximum theoretical speed of transmission on the cable is limited by the requirement that the received signal shall bear a recognizable likeness to the original system of dots and dashes: in other words there is a

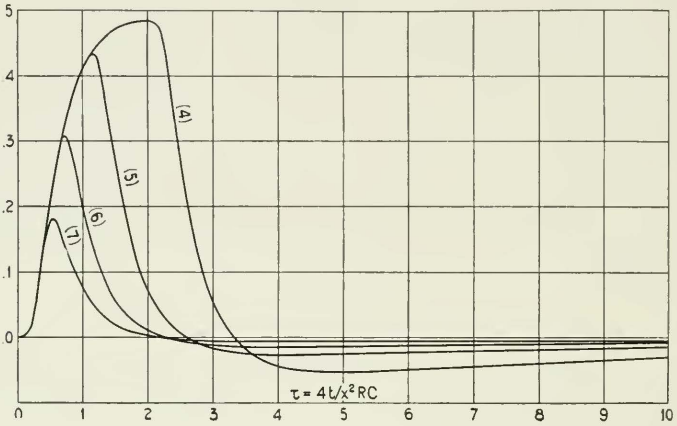


Fig. 6—Elementary telegraph signals in non-inductive cable

maximum allowable departure in wave form between received and transmitted signals. If, therefore, the actual speeds of two cables are inversely proportional to their "KRs," the wave form will be the same. This establishes Kelvin's "KR" law. As a corollary, if the length of the cable is doubled the speed of signaling is reduced to one-quarter, assuming the same definition of signals.

The foregoing will be somewhat clearer, perhaps, if we refer to curves, 4, 5, 6, 7 of Fig. 6 which illustrate the distortion suffered by elementary dot signals in cable transmission. Curve 4 shows the dot signal produced by a unit battery applied to the cable terminals for a time interval $t=2\,\frac{x^2RC}{4}$, while curves 5, 6 and 7 are the corresponding dot signals when the battery is applied for the time intervals $\frac{x^2RC}{4}$, $\frac{1}{2}\,\frac{x^2RC}{4}$ and $\frac{1}{4}\,\frac{x^2RC}{4}$. Any further decrease in the duration of the impressed dot, beyond that shown in curve 7, does not

affect the *shape* of the transmitted dot, which means that the cable speed has reached its theoretical maximum. These curves, it should be observed, can be interpreted in two ways. First, we can regard the length x of the cable as fixed and the duration of the impressed dot as varied. On the other hand, we can regard the actual duration of the impressed dot as constant and the length of the cable as varied. From the latter standpoint the curves illustrate the progressive distortion of the signal as it is transmitted along the cable.

The dot signal of relative duration T can be written as

$$\begin{aligned} D &= I(\tau), \qquad \tau < T \\ &= I(\tau) - I(\tau - T), \qquad \tau < T \end{aligned}$$

and the second expression can be expanded in a Taylor's series, giving

$$D = T\frac{d}{d\tau} I(\tau) - \frac{T^2}{2!}\frac{d^2}{d\tau^2} I(\tau) + \ldots$$

If T is sufficiently short this becomes

$$D = T\ I'(\tau). \tag{173}$$

Hence when the dot signal is of sufficiently short relative duration T, the wave shape of the received signal is constant, $I'(\tau)$, and its amplitude is proportional to the relative duration of the dot.

This can be generalized for any type of transmission system: Let the dot signal be produced by an e.m.f. $f(t)$ of actual duration T. Then the received dot signal, by formula (31), is

$$\begin{aligned} D &= \frac{d}{dt}\int_0^t f(\tau) I(t-\tau) d\tau, \qquad t < T \\ &= \frac{d}{dt}\int_0^T f(\tau) I(t-\tau) d\tau, \qquad t > T. \end{aligned}$$

For $t > T$ this becomes

$$D = I'(t)\int_0^T f(\tau) d\tau - I''(t)\int_0^T \tau f(\tau) d\tau + \ldots$$

and for sufficiently short duration T, we have approximately,

$$D = I'(t)\int_0^T f(\tau) d\tau. \tag{174}$$

Hence for a sufficiently short duration of the impressed e.m.f. the received dot signal is of constant wave form, independent of the shape of the impressed e.m.f., and its amplitude is proportional to the time

integral of the impressed e.m.f. These principles are of considerable practical importance in telegraphy.

The *leaky cable*, that is, a cable with distributed leakage conductance G in addition to resistance R and capacity C, is of some interest. The differential equations of the problem are given in equations (70); the operational formulas for the case of voltage directly impressed on the terminals of the infinitely long line are

$$V = e^{-x\sqrt{CRp+RG}}\ V_o,$$

$$I = \sqrt{\frac{pC}{R} + \frac{G}{R}}\, e^{-x\sqrt{CRp+RG}}\ V_o.$$

Writing $CRx^2 = \alpha$ and $RGx^2 = \beta$, $G/C = \lambda$, and assuming a "unit e.m.f." impressed on the cable, this becomes

$$V = e^{-\sqrt{\alpha p+\beta}}, \tag{175}$$

$$I = \sqrt{\frac{C}{R}}\sqrt{p+\lambda}\ e^{-\sqrt{\alpha p+\beta}}. \tag{176}$$

These equations are readily solved by means of the table and formulas given in a preceding chapter.

But first let us attempt to solve the operational equation (175) for the voltage by Heaviside methods, guided by the solution of the operational equation

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

of the preceding chapter. Expand the exponential function in (175) in the usual power series; it is

$$V = 1 - \sqrt{\alpha p+\beta} + \frac{(\alpha p+\beta)}{2!} - \frac{(\alpha p+\beta)\sqrt{\alpha p+\beta}}{3!} + \cdots \tag{177}$$

Now discard the integral terms and write

$$V = 1 - \left\{1 + \frac{\alpha p+\beta}{3!} + \frac{(\alpha p+\beta)^2}{5!} + \cdots\right\}\sqrt{\alpha p+\beta}. \tag{178}$$

We have now to interpret the expression $\sqrt{\alpha p+\beta}$. We have by ordinary algebra

$$\begin{aligned}\sqrt{\alpha p+\beta} &= \left(1+\frac{\beta}{\alpha p}\right)^{1/2}\sqrt{\alpha p} = \left(1+\frac{\lambda}{p}\right)^{1/2}\sqrt{\alpha p} \\ &= \left[1+\frac{\lambda}{2p} - \frac{1}{2!}\left(\frac{\lambda}{2p}\right)^2 + \frac{1.3}{3!}\left(\frac{\lambda}{2p}\right)^3 + \cdots\right]\sqrt{\alpha p}.\end{aligned} \tag{179}$$

Now identify $\sqrt{p}$ with $1/\sqrt{\pi t}$ in accordance with the Heaviside rule, and $1/p$ with $\int dt$. We get

$$\sqrt{\alpha p+\beta}=\sqrt{\frac{\alpha}{\pi t}}\left\{1+\frac{\lambda t}{1!}-\frac{(\lambda t)^2}{3!}+\frac{1.4}{5!}(\lambda t)^3-\ldots\right\}. \tag{180}$$

Now in the terms of the expansion (178) identify p^n with d^n/dt^n and substitute (180); we get

$$\begin{aligned} V=1-&\left\{1+\frac{1}{3!}\left(\alpha\frac{d}{dt}+\beta\right)+\frac{1}{5!}\left(\alpha^2\frac{d^2}{dt^2}+2\alpha\beta\frac{d}{dt}+\beta^2\right)+\ldots\right\} \\ &\times\sqrt{\frac{\alpha}{\pi t}}\left\{1+\frac{\lambda t}{1!}-\frac{(\lambda t)^2}{3!}+1.4\frac{(\lambda t)^3}{5!}-\ldots\right\}. \end{aligned} \tag{181}$$

This series is hopelessly complicated to either interpret or compute. It is, in fact, an excellent illustration of the grave disadvantages under which many of Heaviside's series solutions labor. We shall therefore attack the solution by aid of the theorems and formulas of a preceding section. The simplicity of the solutions which result is remarkable.

The operational formula for the voltage is

$$V=e^{-\sqrt{\alpha p+\beta}}. \tag{175}$$

Now the operational formula for the voltage in the non-leaky cable is (see equation (164))

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

In order to distinguish between the two cases, let us denote the voltage in the latter case by V^o; thus

$$V^o=e^{-\sqrt{\alpha p}}. \tag{182}$$

Now by theorem (VII) and equation (182) we have

$$\begin{aligned} V^o e^{-\lambda t}&=\frac{p}{p+\lambda}\,e^{-\sqrt{\alpha(p+\lambda)}}, \\ &=\frac{p}{p+\lambda}\,e^{-\sqrt{\alpha p+\beta}}. \end{aligned} \tag{183}$$

Now write (175) as

$$\begin{aligned} V&=\frac{p+\lambda}{p}\cdot\frac{p}{p+\lambda}\,e^{-\sqrt{\alpha p+\beta}}, \\ &=\left(1+\frac{\lambda}{p}\right)\cdot\frac{p}{p+\lambda}\,e^{-\sqrt{\alpha p+\beta}}. \end{aligned} \tag{184}$$

It follows at once by comparison with (183) and the rule that $1/p$ is to be replaced by $\int dt$, that

$$V = \left(1 + \lambda \int_0^t dt\right) V^o e^{-\lambda t}. \tag{185}$$

By a precisely similar procedure with the operational formula (176) for the current, we get

$$I = \left(1 + \lambda \int_0^t dt\right) I^o e^{-\lambda t} \tag{186}$$

where I^o is the current in the non-leaky cable. Now by formulas (169) and (170)

$$V^o = \frac{1}{\sqrt{\pi}} \int_0^{4t/a} \frac{e^{-1/t}}{t\sqrt{t}}\, dt, \tag{169}$$

$$I^o = \sqrt{\frac{C}{\pi R t}}\, e^{-a/4t}, \tag{170}$$

which completes the formal solution of the problem.

Formulas (185) and (186) are extremely interesting, first as showing the superiority of the definite integral to the series expansion—compare (185) with the series expansions (181)—and secondly as exhibiting clearly the effect of leakage on the propagated waves of current and voltage. We see that in both the current and voltage the effect of leakage is two-fold: first it attenuates the wave by the factor $e^{-\lambda t}$, $(\lambda = G/C)$, and secondly it adds a component consisting of the progressive integral of the attenuated wave. This, it may be remarked, is the general effect of leakage in all types of transmission systems. Its effect is, therefore, easily computed and interpreted.

Formulas (185) and (186) are very easy to compute with the aid of a planimeter or integraph; or, failing these devices, by numerical integration. However, for large values of t, the character of the waves is more clearly exhibited if we make use of the identity

$$\int_0^t dt = \int_0^\infty dt - \int_t^\infty dt$$

whence

$$V = \left(1 + \lambda \int_0^\infty dt\right) V^o e^{-\lambda t} - \lambda \int_t^\infty V^o e^{-\lambda t} dt \tag{187}$$

and

$$I = \left(1 + \lambda \int_0^\infty dt\right) I^o e^{-\lambda t} - \lambda \int_t^\infty I^o e^{-\lambda t} dt. \tag{188}$$

The first two terms of these formulas are clearly the ultimate steady state values of the voltage and current waves, and can be determined by evaluating the infinite integrals. A far simpler and more direct way, however, is to make use of the fact that the ultimate steady values of V and I are gotten from the operational formulas by setting $p=0$. That this statement is true is easily seen if we reflect that the steady d.c. voltage and current are gotten from the original differential equations of the problem by assuming a steady state and setting $d/dt=0$.

From the operational formulas we get, therefore,

$$\left(1+\lambda\int_0^{\infty} dt\right) V^o e^{-\lambda t} = e^{-\sqrt{\beta}} = e^{-x\sqrt{RG}}, \tag{189}$$

$$\left(1+\lambda\int_0^{\infty} dt\right) I^o e^{-\lambda t} = \sqrt{\frac{C\lambda}{R}}\, e^{-\sqrt{\beta}} = \sqrt{\frac{G}{R}}\, e^{-x\sqrt{RG}}. \tag{190}$$

Introducing these expressions into (187) and (188) respectively, we get

$$V = e^{-x\sqrt{RG}} - \lambda\int_t^{\infty} V^o e^{-\lambda t} dt, \tag{191}$$

$$I = \sqrt{\frac{G}{R}}\, e^{-x\sqrt{RG}} - \lambda\int_t^{\infty} I^o e^{-\lambda t} dt. \tag{192}$$

The definite integrals can be expanded by partial integration; thus

$$-\lambda\int_t^{\infty} V^o e^{-\lambda t} dt = \int_t^{\infty} V^o de^{-\lambda t}$$

$$= -V^o e^{-\lambda t} - \int_t^{\infty} e^{-\lambda t}\frac{d}{dt} V^o dt.$$

Continuing this process we get

$$V = e^{-x\sqrt{RG}} - e^{-\lambda t}\left(1+\frac{d}{\lambda dt}+\frac{d^2}{\lambda^2 dt^2}+\ \ldots\right) V^o, \tag{193}$$

$$I = \sqrt{\frac{G}{R}}\, e^{-x\sqrt{RG}} - e^{-\lambda t}\left(1+\frac{d}{\lambda dt}+\frac{d^2}{\lambda^2 dt^2}+\ \ldots\right) I^o. \tag{194}$$

Using the values of V^o and I^o, as given by (169) and (170), it is extremely easy to compute V and I, for large values of t, from (193) and (194).

So far we have considered the current and voltage waves in response to a "unit e.m.f.," impressed on the cable at $x=0$. It is of interest and importance to examine the waves due to sinusoidal e.m.fs., suddenly impressed on the cable, particularly in view of proposals to employ alternating currents in cable telegraphy.

We start with the fundamental formula

$$x(t)=\frac{d}{dt}\int_0^t f(t-\tau)h(\tau)d\tau$$
$$=\int_0^t f(t-\tau)h'(\tau)d\tau$$

provided $h(o)=0$, which is the case in the cable.

If $f(t)=\sin \omega t$, we write

$$x_s(t)=\sin \omega t\int_0^t \cos \omega t.h'(t)dt$$
$$-\cos \omega t\int_0^t \sin \omega t.h'(t)dt. \tag{194-a}$$

Similarly, if the impressed e.m.f. is $\cos \omega t$,

$$x_c(t)=\cos \omega t\int_0^t \cos \omega t.h'(t)dt$$
$$+\sin \omega t\int_0^t \sin \omega t.h'(t)dt. \tag{194-b}$$

The investigation of the building-up of alternating currents and voltages, therefore, depends on the progressive integrals

$$C=\int_0^t \cos \omega t.h'(t)dt,$$
$$S=\int_0^t \sin \omega t.h'(t)dt. \tag{194-c}$$

For the case of the *voltage* waves on the non-inductive, non-leaky cable these integrals, by aid of equations (169), become, if we write $\omega'=\alpha\omega/4$,

$$C=\frac{1}{\sqrt{\pi}}\int_0^\tau \frac{e^{-1/\tau}\cos \omega'\tau}{\tau\sqrt{\tau}}d\tau,$$
$$S=\frac{1}{\sqrt{\pi}}\int_0^\tau \frac{e^{-1/\tau}\sin \omega'\tau}{\tau\sqrt{\tau}}d\tau, \tag{194-d}$$

where, as before, $\tau=4t/\alpha$.

For the current wave we have, by (170),

$$C=\frac{2}{\sqrt{\pi}\,xR}\int_0^\tau \left(\frac{1}{\tau}-\frac{1}{2}\right)\frac{e^{-1/\tau}\cos \omega'\tau}{\tau\sqrt{\tau}}d\tau,$$
$$S=\frac{2}{\sqrt{\pi}\,xR}\int_0^\tau \left(\frac{1}{\tau}-\frac{1}{2}\right)\frac{e^{-1/\tau}\sin \omega'\tau}{\tau\sqrt{\tau}}d\tau. \tag{194-e}$$

For small values of τ and ω' these integrals can be numerically evaluated without great labor. Mechanical devices, such as the Coradi Harmonic Analyzer, are here of great assistance. In fact the Coradi Analyzer gives these progressive integrals automatically. It may be said, therefore, that a complete mathematical investigation of the building-up of alternating current and voltage waves on the non-inductive cable presents no serious difficulties, although the labor of computation is necessarily considerable. One fact makes the complete investigations much less laborious than might be sup-

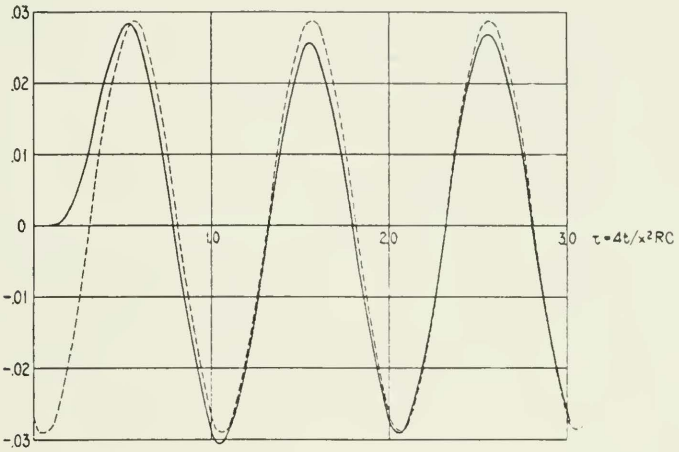


Fig. 7—Non-inductive cable ($G=0$), building-up of alternating current.

Applied e.m.f. $\cos \omega t$; $\omega = 2\pi \frac{4}{x^2RC}$

posed. This is, if the foregoing integrals are calculated for a given value of ω', the results apply to all lengths of cable and all actual frequencies $\omega/2\pi$, such that $\alpha\omega$ is a constant. Then if we double the length of the cable and quarter the frequency, the integrals are unaffected.

The solid curve of Fig. 7 shows the building-up of the cable voltage in response to an e.m.f. $\cos \omega t$, impressed at time $t=0$. The frequency $\omega/2\pi$ is so chosen that $\omega' = \alpha\omega/4 = 2\pi$, and the curve is calculated from equations (194-b) and (194-e). The dotted curve shows the corresponding *steady-state* voltage on the cable; that is, the voltage which would exist if the e.m.f. $\cos \omega t$ had been applied at a long time preceding $t=$''. We observe that, for this frequency, the building-up is effectually accomplished in about one cycle, and that the transient distortion is only appreciable during the first half-cycle.

The case is very much different when a higher frequency is applied. Fig. 8 shows the building-up of the alternating current in the cable when an e.m.f. sin ωt is applied at time $t=0$. The frequency is so chosen that $\omega' = \alpha\omega/4 = 10\pi$. The outstanding features of this curve are that the initial current surge is very large compared with the final steady-state, and that the transient distortion is relatively very large. It is evident that the frequency here shown could not be

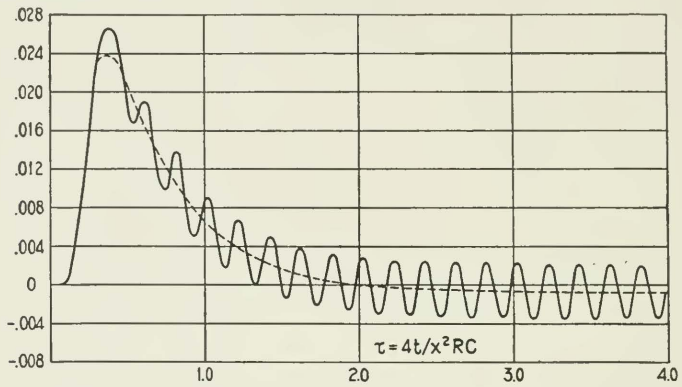


Fig. 8—Non-inductive cable ($G=0$). Building-up of alternating current.

Applied e.m.f. sin ωt; $\omega = 10\pi \frac{4}{x^2RC}$

employed for signaling purposes. This curve has been computed from the steady-state formulas, and equations (160) and (161) for the transient distortion.

If the applied frequency $\omega/2\pi$ is very high, the steady-state becomes negligibly small, and the complete current is obtained to a good approximation by taking the leading terms of (160) and (161). Thus if the applied e.m.f. is sin ωt, and ω is sufficiently large, the cable current is

$$\frac{2}{\sqrt{\pi x R}} \frac{1}{\omega'} \frac{d}{d\tau} \frac{e^{-1/\tau}}{\sqrt{\tau}}$$

by (160) and (170) while, if the impressed e.m.f. is cos ωt, it is

$$\frac{2}{\sqrt{\pi x R}} \left(\frac{1}{\omega'}\right)^2 \frac{d^2}{d\tau^2} \frac{e^{-1/\tau}}{\sqrt{\tau}}$$

by (161) and (170). Here $\omega' = \alpha\omega/4$ and $\tau = 4t/\alpha$.

CHAPTER VII

The Propagation of Current and Voltage Along the Transmission Line

We now take up the more important and difficult problem of investigating the propagation phenomena in the transmission line. The transmission line has distributed series resistance R and inductance L, and distributed shunt capacity C and leakage conductance G. It is the addition of the series inductance L which makes our problem more difficult and at the same time introduces the phenomena of true propagation with finite velocity, as distinguished from the diffusion phenomena of the cable problem. The cable theory serves very well for the problems of trans-oceanic telegraphy [8] but is quite inadequate in the problems of telephonic transmission.

If I denotes the current and V the voltage at point x on the line, the well known differential equations of the problem are:—

$$\left(L\frac{d}{dt}+R\right) I = -\frac{\partial}{\partial x}V,$$
$$\left(C\frac{d}{dt}+G\right) V = -\frac{\partial}{\partial x}I. \tag{195}$$

Replacing d/dt by p, these become

$$(Lp+R)\, I = -\frac{\partial}{\partial x}V,$$
$$(Cp+G)\, V = -\frac{\partial}{\partial x}I. \tag{196}$$

From the second of these equations

$$\frac{\partial V}{\partial x} = -\frac{1}{Cp+G}\frac{\partial^2}{\partial x^2}I$$

and substitution in the first gives

$$(Lp+R)(Cp+G)I = \frac{\partial^2}{\partial x^2}I. \tag{197}$$

Similarly if we eliminate I, we get

$$(Lp+R)(Cp+G)\, V = \frac{\partial^2}{\partial x^2}V. \tag{198}$$

[8] With the installation of the new submarine cable, continuously loaded with permalloy, this statement must be modified. In this cable, the inductance plays a very important part, and is responsible for the greatly increased speed of signaling obtainable.

If we assume a solution of the form

$$V = Ae^{-\gamma x} + Be^{\gamma x}$$

where A and B are arbitrary constants, substitution shows that the solution satisfies the differential equation for V provided

$$\gamma^2 = (Lp+R)(Cp+G). \tag{199}$$

From equation (196) it then follows that

$$\begin{aligned} I &= \frac{\gamma}{Lp+R}(Ae^{-\gamma x} - Be^{\gamma x}) \\ &= \frac{Cp+G}{\gamma}(Ae^{-\gamma x} - Be^{\gamma x}). \end{aligned} \tag{200}$$

Now restricting attention to the infinitely long line extending along the positive x axis, with voltage V_o impressed directly on the line at $x=0$, the reflected wave vanishes and we get

$$\begin{aligned} V &= V_o e^{-\gamma x}, \\ I &= \frac{Cp+G}{\gamma} V_o e^{-\gamma x}, \\ \gamma^2 &= (Lp+R)(Cp+G). \end{aligned} \tag{201}$$

Now let us write

$$\gamma^2 = \frac{1}{v^2}[(p+\rho)^2 - \sigma^2] \tag{202}$$

where

$$\begin{aligned} v &= 1/\sqrt{LC}, \\ \rho &= \frac{R}{2L} + \frac{G}{2C}, \\ \sigma &= \frac{R}{2L} - \frac{G}{2C}. \end{aligned}$$

Then setting $V_o = 1$, the *operational equations* of the problem become

$$V = e^{-\frac{x}{v}\sqrt{(p+\rho)^2-\sigma^2}}, \tag{203}$$

$$I = v\left(C + \frac{G}{p}\right) p \frac{e^{-\frac{x}{v}\sqrt{(p+\rho)^2-\sigma^2}}}{\sqrt{(p+\rho)^2-\sigma^2}}. \tag{204}$$

Now consider the operational equation, defining a new variable F:

$$F = p \frac{e^{-\frac{x}{v}\sqrt{(p+\rho)^2-\sigma^2}}}{\sqrt{(p+\rho)^2-\sigma^2}}. \tag{205}$$

It follows at once from our operational rules, and (203) and (204), that

$$I = v\left(C + G\int_0^t dt\right)F, \tag{206}$$

$$V = -v\int_0^t \frac{\partial F}{\partial x}\,dt. \tag{207}$$

Our problem is thus reduced to evaluating the function F, from the operational equation (205). This equation can be solved by aid of the operational rules and formulas already given. The process is rather complicated, and there is less chance of error if we deal instead with the integral equation of the problem

$$\frac{e^{-\frac{x}{v}\sqrt{(p+\rho)^2-\sigma^2}}}{\sqrt{(p+\rho)^2-\sigma^2}} = \int_0^\infty F(t)e^{-pt}dt. \tag{208}$$

Now let us search through our table of definite integrals. We do not find this integral equation as it stands, but we do observe that formula (m) resembles it, and this resemblance suggests that formula (m) can be suitably transformed to give the solution of (208). We therefore start with the formula

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

This, regarded as an integral equation, defines a function which is zero for $t<\lambda$ and has the value $J_o(\sqrt{t^2-\lambda^2})$ for $t\geqq\lambda$, J_o being the Bessel function of order zero. We now transform (m) as follows:

(1) Let $\lambda p = q$ and $t/\lambda = t_1$. Substituting in (m) we get

$$\frac{e^{-\sqrt{q^2+\lambda^2}}}{\sqrt{q^2+\lambda^2}} = \int_1^\infty e^{-qt_1}J_o(\lambda\sqrt{t_1^2-1})\,dt_1.$$

Now, in order to keep our original notation in p and t, replace q by p and t_1 by t; we get

$$\frac{e^{-\sqrt{p^2+\lambda^2}}}{\sqrt{p^2+\lambda^2}} = \int_1^\infty e^{-pt}J_o(\lambda\sqrt{t^2-1})\,dt. \tag{m.1}$$

(2) In (m.1) make the substitution $p = q+\mu$ and then in the final expression replace q by p; we get

$$\int_1^\infty e^{-pt}.e^{-\mu t}J_o(\lambda\sqrt{t^2-1})dt = \frac{e^{-\sqrt{(p+\mu)^2+\lambda^2}}}{\sqrt{(p+\mu)^2+\lambda^2}}. \tag{m.2}$$

(3) In (m.2) make the substitution $p=\frac{x}{v}q$ and $t_2=\frac{x}{v}t$, and ultimately replace q by p and t_2 by t; we get

$$\int_{x/v}^{\infty} e^{-pt}.e^{-\mu_1 t}J_o\left(\lambda_1\sqrt{t^2-\frac{x^2}{v^2}}\right)dt=\frac{e^{-\frac{x}{v}\sqrt{(p+\mu_1)^2+\lambda_1^2}}}{\sqrt{(p+\mu_1)^2+\lambda_1^2}} \tag{m.3}$$

where $\lambda_1=\frac{v}{x}\lambda$ and $\mu_1=\frac{v}{x}\mu$. (They are, of course, as yet, arbitrary parameters, except that they are restricted to positive values).

(4) Now if we compare (m.3) with the integral equation (208) for F, we see that they are identical provided we get

$$\mu_1=\rho,$$
$$\lambda_1=i\sigma=\sigma\sqrt{-1},$$

which is possible, since $\rho>\sigma$.

Introducing these relations, we have

$$\int_{x/v}^{\infty} e^{-pt}.e^{-\rho t}\ I_o(\sigma\sqrt{t^2-x^2/v^2})dt=\frac{e^{-\frac{x}{v}\sqrt{(p+\rho)^2-\sigma^2}}}{\sqrt{(p+\rho)^2-\sigma^2}}. \tag{m.4}$$

Here I_o denotes the Bessel function of imaginary argument; thus $J_o(iz)=I_o(z)$.

It follows from (m.4) and the integral equation (208) that

$$\begin{aligned} F(t)&=0 \text{ for } t<x/v, \\ &=e^{-\rho t}I_o(\sigma\sqrt{t^2-x^2/v^2}) \text{ for } t\geqq x/v. \end{aligned} \tag{209}$$

Having now solved for $F=F(t)$, the current and voltage are gotten from equations (206) and (207). Thus

$$\begin{aligned} I&=0 \text{ for } t<x/v, \\ &=\sqrt{\frac{C}{L}}\,F(t)+vG\int_{x/v}^{t}F(t)dt \text{ for } t\geqq x/v. \end{aligned} \tag{210}$$

The corresponding voltage formula is

$$\begin{aligned} V&=0 \text{ for } t<x/v, \\ &=e^{-\rho x/v}+\frac{\sigma x}{v}\int_{x/v}^{t}\frac{e^{-\rho\tau}I_1(\sigma\sqrt{\tau^2-x^2/v^2})}{\sqrt{\tau^2-x^2/v^2}}d\tau \text{ for } t\geqq x/v. \end{aligned} \tag{211}$$

Here $I_1(\sigma\sqrt{\tau^2-x^2/v^2})$ is the Bessel function of order 1: thus $-iJ_1(iz)=I_1(z)$. The function is entirely real. The derivation of formula (211) is a little troublesome, owing to the discontinuous character of the function F: the detailed steps are given in an appendix.

The preceding solution depends for its outstanding directness and simplicity on the recognition of the infinite integral identity (m), into which the integral equation of the problem can be transformed. When such identities are known their value in connection with the solution of operational equations requires no emphasis. On the other hand, we cannot always expect to find such an identity in the case of every operational equation; and, particularly in the case of such an important case as the transmission equation it would be unfortunate to have no alternative mode of solution. Fortunately a quite direct series expansion solution is obtainable from the operational equation, and this will now be derived. As a matter of convenience we shall restrict the derivation to the voltage formula

$$V = e^{-\frac{x}{v}\sqrt{(p+\rho)^2-\sigma^2}}. \tag{203}$$

As a further matter of mere convenience we shall assume that $G=0$, so that $\sigma=\rho$ and (203) becomes

$$V = e^{-\tau\sqrt{p^2+2\rho p}} \tag{203-a}$$

where $\tau = x/v$.

The method holds equally well for the current equation (204) and for the general case $\sigma \neq \rho$.

Write (203-a) as

$$V = e^{-\tau p(1+2\rho/p)^{1/2}}$$

and expand the exponential factor $(1+2\rho/p)^{1/2}$ by the binomial theorem; thus

$$(1+2\rho/p)^{1/2} = 1+\frac{\rho}{p}+\alpha_2\left(\frac{\rho}{p}\right)^2+\alpha_3\left(\frac{\rho}{p}\right)^3+\ldots$$

so that

$$V = e^{-\tau p}.e^{-\rho\tau}.\exp\left(-\frac{\alpha_2\tau\rho^2}{p}-\frac{\alpha_3\tau\rho^3}{p^2}-\frac{\alpha_4\tau\rho^4}{p^3}-\ldots\right).$$

Now the operational equation

$$v = \exp\left(-\frac{\alpha_2\tau\rho^2}{p}-\frac{\alpha_3\tau\rho^3}{p^2}-\frac{\alpha_4\tau\rho^4}{p^3}-\ldots\right)$$

can be expanded in inverse powers of p; thus

$$v = 1+\frac{\beta_1}{p}+\frac{\beta_2}{p^2}+\frac{\beta_3}{p^3}+\ldots\ldots$$

the power series solution of which is

$$v(t) = 1+\frac{\beta_1 t}{1!}+\frac{\beta_2 t^2}{2!}+\frac{\beta_3 t^3}{3!}+\ldots\ldots$$

It follows at once from the preceding and Theorem VII that

$$V(t)=0 \text{ for } t<\tau$$
$$=e^{\rho\tau}\left(1+\beta_1\frac{(t-\tau)}{1!}+\beta_2\frac{(t-\tau)^2}{2!}+\ldots.\right) \text{ for } t>\tau$$

If the coefficients β_1, β_2 . . . are evaluated, a simple matter of elementary algebra, the foregoing expansion in the retarded time $t-\tau$ will be found to agree with the solution (211) when σ is put equal to ρ.

We shall now discuss the outstanding features of the propagation phenomena in the light of equations (210) and (211) for the current and voltage. We observe, first, that we have a true finite velocity of propagation $v=1/\sqrt{LC}$. No matter what the form of impressed e.m.f. at the beginning of the line ($x=0$), its effect does not reach the point x of the line until a time $t=x/v$ has elapsed. Consequently $v=x/t$ is the velocity with which the wave is propagated. This is a strict consequence of the distributed inductance and capacity of the line and depends only on them, since $v=1/\sqrt{LC}$. It will be recalled that in the case of the cable, where the inductance is ignored, no finite velocity of propagation exists.

The question of velocity of propagation of the wave has been the subject of considerable confusion and misinterpretation when dealing with the steady-state phenomena. It seems worth while to briefly touch on this in passing.

As has been pointed out in preceding chapters, the symbolic or complex steady-state formula is gotten from the operational equation by replacing the symbol p by $i\omega$ where $i=\sqrt{-1}$ and $\omega/2\pi$ is the frequency. If this is done in the operational equation (203) for the voltage, the symbolic formula is

$$V=e^{\frac{x}{v}\sqrt{(i\omega+\rho)^2-\sigma^2}}e^{i\omega t}.$$

If the expression $\sqrt{(i\omega+\rho)^2-\sigma^2}$ is separated into its real and imaginary parts we get an expression of the form

$$V=e^{-\alpha x}e^{i\omega\left(t-\frac{\beta x}{v}\right)},$$

where

$$\beta=\sqrt{\frac{\omega^2+\sigma^2-\rho^2+\sqrt{(\omega^2+\sigma^2-\rho^2)^2+4\omega^2\rho^2}}{2\omega^2}}$$

and

$$\alpha=\rho/\beta v.$$

Now if we keep the expression $t-\beta\frac{x}{v}$ constant, that is, if we move along the line with velocity $dx/dt=v/\beta$, the phase of the wave will remain constant. This is interpreted often as meaning that the

velocity of propagation of the wave is v/β. Now since β is greater than unity and only approaches unity as the frequency becomes indefinitely great, the inference is frequently made that the velocity of propagation depends upon and increases to a limiting value v, with the frequency. This velocity, however, is not the true velocity of propagation, which is always v, but is the *velocity of phase propagation in the steady-state*. This distinction is quite important and failure to bear it in mind has led to serious mistakes.

Returning to equation (211) and (210) we see that after a time interval $t=x/v$ has elapsed since the unit e.m.f. was impressed on the cable, the voltage at point x suddenly jumps from zero to the value $e^{-\rho x/v}$ while the current correspondingly jumps to the value $\sqrt{\frac{C}{L}}\, e^{-\rho x/v}$. The exponential factor $\rho x/v$ is

$$x\left(\frac{R}{2L}+\frac{G}{2C}\right)\sqrt{LC}=x\left(\frac{R}{2}\sqrt{\frac{C}{L}}+\frac{G}{2}\sqrt{\frac{L}{C}}\right)=ax$$

which will be recognized as the *steady-state attenuation factor* for high frequencies. Similarly $\sqrt{C/L}$ is the steady-state admittance of the line for high frequencies. The sudden jumps in the current and voltage at time $t=x/v$ are called the heads of the current and voltage waves. If, instead of a unit e.m.f., a voltage $f(t)$ is impressed on the line at time $t=0$, the corresponding heads of the waves are $f(o)e^{-ax}$ and $\sqrt{C/L}\ f(o)e^{-ax}$ for voltage and current respectively. These expressions follow at once from the integral formula

$$\begin{aligned} x(t) &= \frac{d}{dt}\int_0^t f(t-\tau)h(\tau)d\tau \\ &= f(o)h(t)+\int_0^t f'(t-\tau)h(\tau)d\tau. \end{aligned}$$

The tails of the waves, that is, the parts of the waves subsequent to the time $t=x/v$, are more complicated and will depend on the distance x along the line and on the line parameters ρ and σ. The two simplest cases are the *non-dissipative* line, and the *distortionless* line.

The ideal non-dissipative line, quite unrealizable in practice, is one in which both R and G are zero. In this case $\rho=\sigma=0$, and formulas (210) and (211) become

$$\begin{aligned} I &= 0 \text{ for } t<x/v, \\ &= \sqrt{\frac{C}{L}} \text{ for } t\geq x/v, \\ V &= 0 \text{ for } t<x/v, \\ &= 1 \text{ for } t\geq x/v. \end{aligned}$$

Both current and voltage jump, at time $t=x/v$, to their steady values. If an e.m.f. $f(t)$ is impressed on the line at time $t=0$, the corresponding current and voltage waves are

$$\begin{aligned} I &= 0 \text{ for } t<x/v, \\ &= \sqrt{\frac{C}{L}} f(t-x/v) \text{ for } t \geq x/v, \\ V &= 0 \text{ for } t<x/v, \\ &= f(t-x/v) \text{ for } t \geq x/v. \end{aligned}$$

Consequently the ideal non-dissipative line transmits the waves with finite velocity v, without attenuation or distortion. Such a line is, of course, the ideal transmission system.

The non-dissipative line is, of course, purely theoretical and unrealizable in practice; the *distortionless* line is, however, approximately realizable, and as the name implies, transmits without distortion of wave form. The distortionless line is one in which the line constants are so related that

$$\sigma = \frac{R}{2L} - \frac{G}{2C} = 0.$$

If this condition is satisfied, formulas (210) and (211) become

$$\begin{aligned} I &= 0 \text{ for } t<x/v, \\ &= \sqrt{\frac{C}{L}} e^{-ax} \text{ for } t \geq x/v, \\ V &= 0 \text{ for } t<x/v, \\ &= e^{-ax} \text{ for } t \geq x/v. \end{aligned}$$

Furthermore, if the impressed e.m.f. is $f(t)$, the corresponding current and voltage waves are:—

$$\begin{aligned} I &= 0 \text{ for } t<x/v, \\ &= \sqrt{\frac{C}{L}} e^{-ax} f(t-x/v) \text{ for } t \geq x/v, \\ V &= 0 \text{ for } t<x/v, \\ &= e^{-ax} f(t-x/v) \text{ for } t \geq x/v. \end{aligned}$$

The distortionless line, therefore, transmits the waves without distortion of wave form, but attenuates the waves by the factor e^{-ax}. Such a line is an ideal transmission system as regards preservation of wave form, but introduces serious attenuation losses. For example, if a line has normally negligible leakage, and leakage is introduced

to secure the condition $R/L=G/C$, the line is thereby rendered distortionless but the attenuation is doubled.

One of Heaviside's most important contributions to wire transmission theory was to point out the properties of the distortionless line, its approximately realizable character, and to base on it a correct theory of telephonic transmission.

The character of the wave propagation when the parameters ρ and σ are not restricted to special values, can only be roughly inferred from inspection of the formulas, and then only when the properties of the Bessel function I_o and I_1 have been studied. Fortunately these functions have been computed and tabulated for small values of the argument, and have simple asymptotic expansions for large values. It is therefore a simple matter to compute and graph a representative set of curves which show the current and voltage waves for various values of ρ, σ and x. For this purpose it is convenient to introduce a change of variables and write:

$$\tau = vt$$
$$a = \rho/v$$
$$b = \sigma/v$$

whence the formulas for current and voltage become:

$$I = \sqrt{\frac{C}{L}}\, e^{-a\tau} I_o(b\sqrt{\tau^2-x^2}) + (a-b)\sqrt{\frac{C}{L}} \int_x^\tau e^{-a\tau} I_o(b\sqrt{\tau^2-x^2})\, d\tau, \tag{210a}$$

$$V = e^{-ax} + bx \int_x^\tau \frac{e^{-a\tau} I_1(b\sqrt{\tau^2-x^2})}{\sqrt{\tau^2-x^2}}\, d\tau. \tag{211a}$$

Figs. (9) to (18) give a representative set of curves illustrating the form of the propagated current and voltage waves for different lengths of line, and different values of the line parameters a and b, or ρ and σ.

The curves of Figs. (9) and (10) show the current entering the line in response to a unit e.m.f. applied at time $t=0$. The line is assumed to be non-leaky ($b=0$) and is computed for two different values of the parameter a. We see that the current instantly jumps to the value $\sqrt{C/L}$ and then begins to die away, the rate at which it dies away depending on and increasing with the parameter $a = \frac{R}{2}\sqrt{\frac{C}{L}}$.

If we now consider a point x out on the line, the current is zero until $\tau=x$, at which time it jumps to the value $\sqrt{C/L}\, e^{-ax}$. It then

begins to die away provided x and a are such that $ax<2$. If, however, we are considering a point at which $ax>2$, the current begins to rise instead of fall after the initial jump, and may attain a maximum value very large compared with the head before it starts to die away. This is shown in the curves of Figs. (11), (12) and (13), also computed

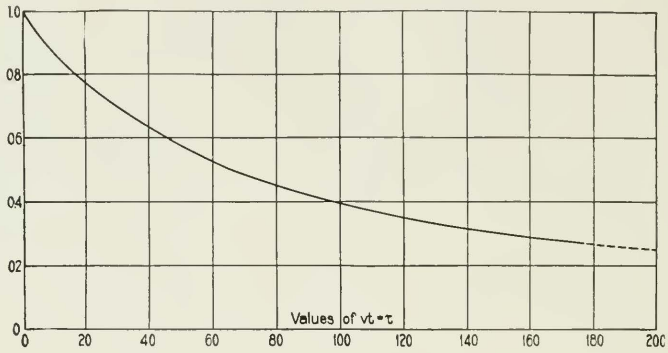


Fig. 9—Current entering line; $\frac{R}{2}\sqrt{\frac{C}{L}}=a=0.0132$; $G=0$.

Multiply ordinates by $\sqrt{C/L}$

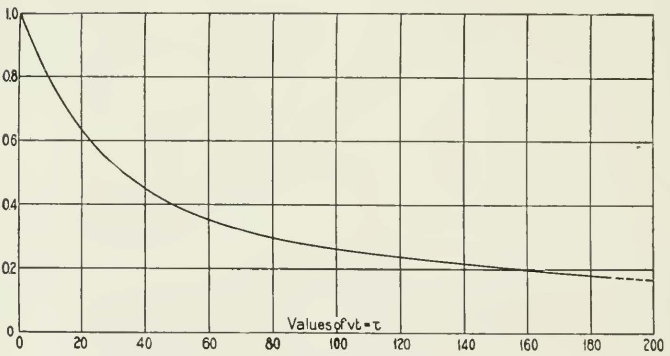


Fig. 10—Current entering line; $\frac{R}{2}\sqrt{\frac{C}{L}}=a=0.2645$; $G=0$.

Multiply ordinates by $\sqrt{C/L}$

for the non-leaky line ($b=0$). From these curves we see that, as the length of the line and the parameter a increase, the relative magnitude of the tail, as compared with the head of the wave, increases. Finally when the line becomes very long, the head of the wave becomes negligibly small, and the wave, except in the neighborhood of its head, becomes very close to that of the corresponding non-inductive

cable. This is shown in curves (13) and (14), for the line and the corresponding cable, which are plotted to the same time scale and ordinate scale to facilitate comparison. Curve (15) shows the effect of leakage in eliminating the tail. This line is not quite distortionless but nearly so.

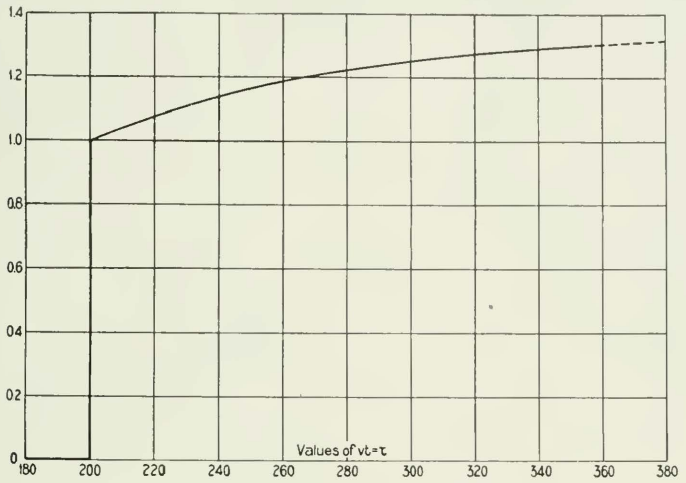


Fig. 11—Propagated current in line; $x = 200$; $\frac{R}{2}\sqrt{\frac{C}{L}} = a = 0.0132$; $G = 0$.

Multiply ordinates by $\sqrt{C/L}.e^{-2.64}$

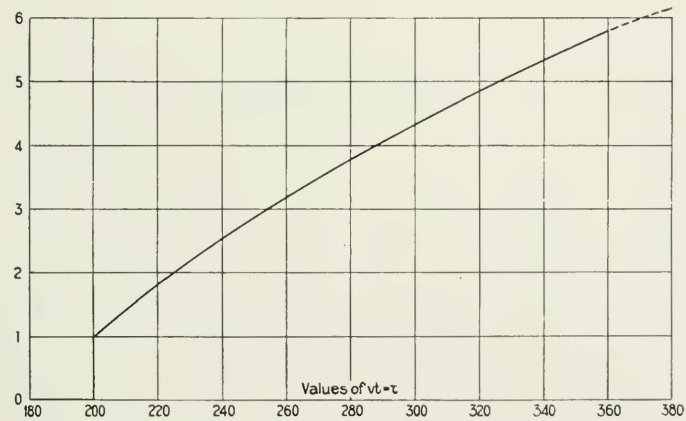


Fig. 12—Propagated current in line; $x = 200$; $\frac{R}{2}\sqrt{\frac{C}{L}} = 0.02645$; $G = 0$.

Multiply ordinates by $\sqrt{C/L}.e^{-5.29}$

An interesting feature of both current and voltage waves is that when a sufficient time has elapsed after the arrival of the head of the wave, the waves become closer and closer to the wave of the corresponding non-inductive cable; that is, to the cable having the same R,C and G. Consequently the inductance plays no part in the subsidence of the waves to their final values.

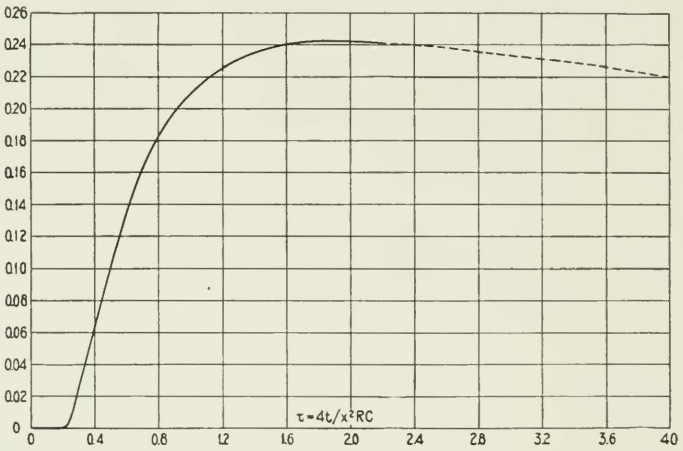


Fig. 13—Propagated current in line; $\frac{R}{2}\sqrt{\frac{C}{L}}x=10$; $G=0$.

Multiply ordinates by $2/Rx$.

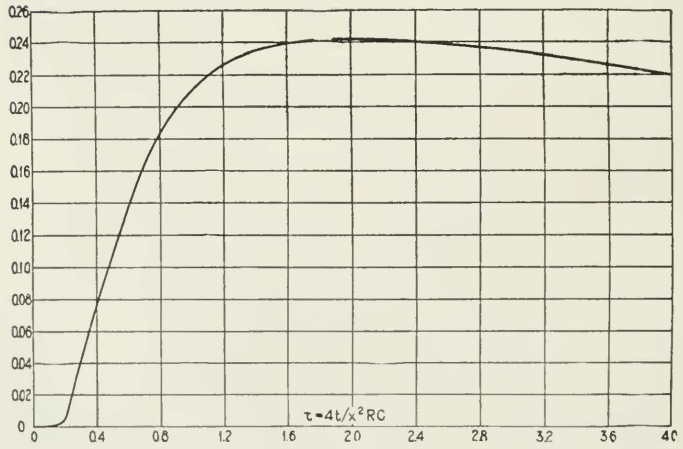


Fig. 14—Propagated current in cable. Multiply ordinates by $2/Rx$.

Curves (16), (17) and (18) illustrate the voltage wave for several conditions. After the arrival of the head, the wave slowly builds up to its final value. Curve (18) represents the case where the line is very nearly distortionless, showing how completely the distorting tail of the wave is eliminated.

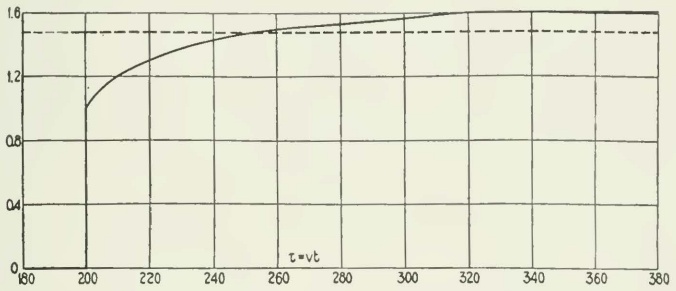


Fig. 15—Propagated current in line; $x=200$

$$a=\frac{R}{2}\sqrt{\frac{C}{L}}+\frac{G}{2}\sqrt{\frac{L}{C}}=0.0353$$

$$b=\frac{R}{2}\sqrt{\frac{C}{L}}-\frac{G}{2}\sqrt{\frac{L}{C}}=0.01765$$

Multiply ordinates by $\sqrt{C/L}.e^{-7.06}$



Fig. 16—Propagated voltage in line; $\frac{R}{2}\sqrt{\frac{C}{L}}\,x=ax=3$; $G=0$.

So far we have confined attention to the current and voltage waves in response to a unit e.m.f. applied at time $t=0$ to the line terminals. Of much greater technical importance is the question of the waves in response to a sinusoidal e.m.f. suddenly applied to the line termi-

nals. In order to investigate this important problem it is convenient to divide the expressions for the current and voltage waves as given by equations (210-a) and (211-a) into two components. We write for $\tau \geq x$,

$$I = \sqrt{\frac{C}{L}} e^{-ax} + J(t), \tag{210-b}$$

$$V = e^{-ax} + W(t), \tag{211-b}$$

where, by definition, $J(t)$ and $W(t)$ are the differences between the total waves and their heads. The advantage of analyzing the waves into these components is that the distortion of the waves is due to

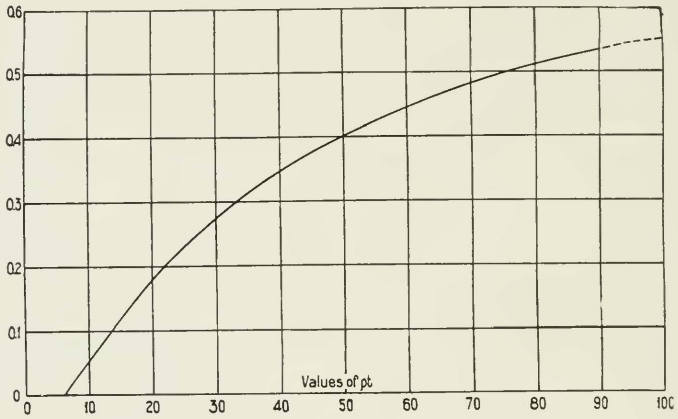


Fig. 17—Propagated voltage in line; $\frac{R}{2}\sqrt{\frac{C}{L}}\, x = ax = 6$; $G = 0$.

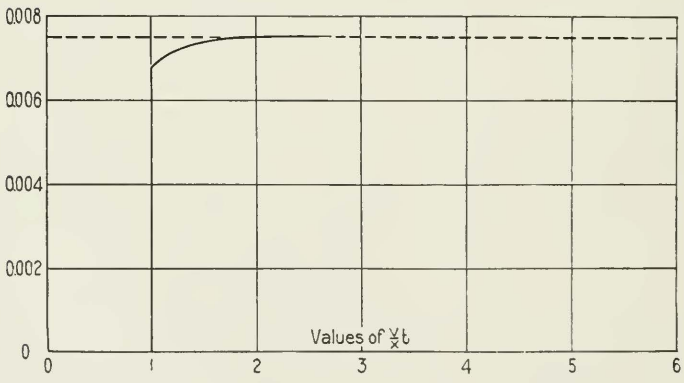


Fig. 18—Propagated voltage in line; $\frac{R}{2}\sqrt{\frac{C}{L}}\, x = ax = 3$; $\frac{G}{2}\sqrt{\frac{L}{C}}\, x = bx = 2$.

$J(t)$ and $W(t)$ respectively, while the first component of (210-b) and (211-b) introduce merely a delay. Thus, if the e.m.f. impressed at time $t=0$ is $f(t)$, the corresponding waves for $t \geq x/v$ or $\tau \geq x$, are

$$I = \sqrt{\frac{C}{L}} e^{-ax} f(t-x/v) + \int_{x/v}^{t} f(t-t_1) J'(t_1) dt_1, \tag{212}$$

$$V = e^{-ax} f(t-x/v) + \int_{x/v}^{t} f(t-t_1) W'(t_1) dt_1, \tag{213}$$

where $J'(t) = \frac{d}{dt} J(t)$ and $W'(t) = \frac{d}{dt} W(t)$.

The integrals of (212) and (213) can be computed and analyzed in precisely the same way as discussed in connection with the non-inductive cable problem, and are of very much the same character as the alternating current waves of the cable. In the total waves, however, as given by (212) and (213), a very essential difference is introduced by the absence of the first terms, which represent undistorted waves propagated with velocity v. Thus, if the impressed e.m.f. is sin ωt, (212) and (213) become

$$I = \sqrt{\frac{C}{L}} e^{-ax} \sin \omega(t-x/v) + \int_{x/v}^{t} \sin \omega(t-t_1) J'(t_1) dt_1, \text{ for } t \geq x/v \tag{214}$$

$$V = e^{-ax} \sin \omega(t-x/v) + \int_{x/v}^{t} \sin \omega(t-t_1) W'(t_1) dt_1, \text{ for } t \geq x/v. \tag{215}$$

Now the first terms of (214) and (215) are simply the usual steady-state expressions for the current and voltage waves when the frequency is sufficiently high to make the steady-state attenuation constant equal to a and the phase velocity equal to v. Furthermore the integral terms become smaller and smaller as the applied frequency $\omega/2\pi$ is increased. It follows, therefore, that for high frequencies the waves assume substantially their final steady value at time $t=x/v$, and that the tails of the waves, or the transient distortion, becomes negligible. This is a consequence entirely of the

presence of inductance in the line, and shows its extreme importance in the propagation of alternating waves and the reduction of transient distortion.

It should be pointed out, however, that if the line is very long and the attenuation is very high, the integral terms of (214) and (215) are not negligible unless the applied frequency is correspondingly very high. For example, on a long submarine cable, the a.c. attenuation is so large that the first terms of (214) and (215) are very small, and $J(t)$ is very large compared with $\sqrt{C/L}\, e^{-ax}$. Consequently here there is very serious transient distortion and alternating currents are therefore not adapted for submarine telegraph signalling.

This discussion may possibly be made a little clearer, without detailed analysis, if we recall the discussion of alternating current propagation in the non-inductive cable of the preceding chapter. From that analysis it follows that, when the applied frequency $\omega/2\pi$ is sufficiently high, the integral term of (214) becomes approximately

$$\frac{1}{\omega} J'(t)$$

and the complete current wave is

$$\sqrt{\frac{C}{L}} e^{-ax} \sin \omega(t-x/v) + \frac{1}{\omega} J'(t) \tag{216}$$

and similarly the voltage wave is

$$e^{-ax} \sin \omega(t-x/v) + \frac{1}{\omega} W'(t). \tag{217}$$

Now if the total attenuation ax is large the last terms of (216) and (217), before they ultimately die away, may become very large compared with the first terms, which represent the ultimate steady-state.

Appendix to Chapter VII. Derivation of Formula (211)

The only troublesome question involved in deriving (211) from (207) and (209) is that we have to differentiate with respect to x, in accordance with (207), the discontinuous function $F(t)$. To accomplish this we write (209) in the form

$$F(t) = \phi(t-x/v) e^{-\rho t} I_o(\sigma \sqrt{t^2 - x^2/v^2}) \tag{209-a}$$

where $\phi(t)$ is defined as a function which is zero for $t < x/v$ and unity for $t \geq x/v$. Clearly this is equivalent to (209) and permits us to deal

with $F(t)$ as a *continuous* function. Now, in accordance with (207), perform the operation of differentiation upon (209-a): we get

$$-v\frac{\partial F}{\partial x} = \frac{\partial}{\partial t}\phi(t-x/v)e^{-\rho t}I_o(\sigma\sqrt{t^2-x^2/v^2})$$

$$-v\phi(t-x/v)\frac{\partial}{\partial x}e^{-\rho t}I_o(\sigma\sqrt{t^2-x^2/v^2}).$$

The first expression follows from the fact that

$$\frac{\partial}{\partial x}\phi(t-x/v) = -\frac{1}{v}\frac{\partial}{\partial t}\phi(t-x/v).$$

We observe also that $\frac{\partial}{\partial t}\phi(t-x/v)=0$ except at $t=x/v$, when it is infinite. We also observe that, for $t\geq x/v$,

$$\int_0^t \frac{\partial}{\partial t}\phi(t-x/v)dt = 1$$

and that the whole contribution to the integral occurs at $t=x/v$. With these points clearly in mind, the expression

$$V = -v\int_0^t \frac{\partial F}{\partial x}dt$$

reduces to (211) without difficulty.

CHAPTER VIII

Propagation of Current and Voltage in Artificial Lines and Wave Filters

The artificial line here considered is a periodic structure, composed of a series of sections connected in tandem, each section consisting of a lumped impedance z_1 in series with the line, and a lumped

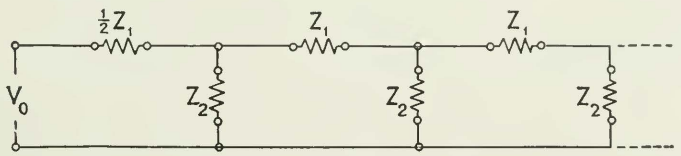


Fig. 19

impedance z_2 in shunt across the line. In the artificial line which we shall consider it will be assumed that the voltage is applied at the middle of the initial or zeroth section, as shown in Fig. 19. This termination is chosen because of its practical importance, and be-

cause also of the fact that the mathematical analysis is simplified thereby. Furthermore any other termination can be regarded and dealt with as an additional terminal impedance, so there is no essential loss of generality involved.

A study of the properties of the artificial line is of practical importance for several reasons:

1. The artificial line is often used as a model of an actual transmission line and it is therefore of importance to determine theoretically the degree of correspondence between the two.

2. The solution for the corresponding transmission line with continuously distributed constants is derivable from the solution for the artificial line by keeping the total inductance, resistance, capacity and leakage constant or finite, and letting the number of sections approach infinity.

3. The artificial line is very closely related, in its properties and performance, to the periodically loaded line, and its solution is, to a first approximation, a working solution for the loaded line.

4. The structure is of great importance in its own right, and when the impedance elements are properly chosen, constitutes a "wave filter."

We shall now derive the operational and symbolic equations which formulate the propagation phenomena in the artificial line. Let I_n denote the mesh current in the nth section of the line; I_{n-1} the mesh current in the $(n-1)^{\text{th}}$ section, etc. Now write down the expression for the voltage drop in the n^{th} section; in accordance with Kirchhoff's law we get:

$$(z_1+2z_2)I_n-z_2(I_{n-1}+I_{n+1})=0 \tag{218}$$

where, of course, the impedances have the usual significance.

Now this is a difference equation, as distinguished from a differential equation, but the method of solution is essentially the same. We *assume* a solution of the form

$$I_n=Ae^{-n\Gamma}+Be^{n\Gamma} \tag{219}$$

where A, B and Γ are independent of n, and substitute in (218). After some simple rearrangements we get

$$\{(z_1+2z_2)-2z_2\cosh\Gamma\}\cdot\{Ae^{-n\Gamma}+Be^{n\Gamma}\}=0. \tag{220}$$

Equation (218) is clearly satisfied by the assumed form of solution, and furthermore leaves the constants A and B arbitrary and at our

disposal to satisfy any boundary conditions, provided Γ is so chosen that

$$\cosh \Gamma = \frac{z_1+2z_2}{2z_2} = 1+2\rho \tag{221}$$

where $\rho = z_1/4z_2$.

Now by reference to equation (219) it is easily seen that Γ is the *propagation constant* of the artificial line, precisely analogous to the propagation constant γ of the smooth line. In terms of the impedances z_1 and z_2, the propagation constant of the artificial line is determined by (221). This equation may either be regarded as an operational equation or a symbolic equation, depending on whether the impedances are expressed in terms of the operator p or in terms of $i\omega$, where ω is 2π times the frequency.

Now suppose in (221) we write $e^{\Gamma} = x$; the equation becomes

$$x + 1/x = 2(1+2\rho)$$

and solving for x we get

$$\begin{aligned} x = e^{\Gamma} &= (1+2\rho) + \sqrt{(1+2\rho)^2 - 1} \\ &= (\sqrt{1+\rho} + \sqrt{\rho})^2 = (\sqrt{1+\rho} - \sqrt{\rho})^{-2} \end{aligned} \tag{222}$$

which is an explicit formula for Γ.

Now return to equation (219) and let us assume that the line is either infinitely long, or, what amounts to the same thing, that it is closed by an impedance which suppresses the reflected wave. We assume also that a voltage V_o is impressed at mid-series position of the zeroth section ($n = $). Equation (219) becomes

$$I_n = A e^{-n\Gamma}$$

and the currents in the zeroth and 1st sections are

$$I_o = A, \qquad I_1 = A e^{-\Gamma}.$$

Now, by direct application of Kirchhoff's law to the zeroth section, we have

$$V_o = (\tfrac{1}{2}z_1 + z_2) I_o - z_2 I_1,$$

whence

$$A\{\tfrac{1}{2}z_1 + z_2(1-e^{-\Gamma})\} = V_o. \tag{223}$$

But

$$I_o = A = \frac{1}{K} V_o,$$

$$I_n = \frac{V_o}{K} e^{-n\Gamma},$$

where K is the *characteristic impedance* of the artificial line (at mid-series position). Hence by (223) and (222)

$$\frac{1}{K}=\frac{1}{z_2}\frac{1}{(1-e^{-\Gamma})+2\rho}$$
$$=\frac{1}{2z_2}\frac{1}{\sqrt{\rho+\rho^2}}=\frac{1}{\sqrt{z_1z_2}}\frac{1}{\sqrt{1+\rho}}. \tag{224}$$

By aid of the preceding the direct current wave can be written as

$$I_n=\frac{V_o}{\sqrt{z_1z_2}}\frac{[\sqrt{1+\rho}-\sqrt{\rho}\,]^{2n}}{\sqrt{1+\rho}}. \tag{225}$$

This formula is not so physically suggestive as its equivalent

$$I_n=\frac{V_o}{K}e^{-n\Gamma}$$

but is useful when we come to the solution of the operational equation.

Before proceeding with the operational equation, and the investigation of transient phenomena in artificial lines, it will be of interest to deduce from the foregoing the unique and remarkable properties of wave filters in the steady state. For this purpose we return to equation (221)

$$\cosh\,\Gamma=1+2\rho.$$

Now suppose that the series impedance z_1 is an inductance L and the shunt impedance z_2 a capacity C, so that, symbolically,

$$z_1=i\omega L,\ z_2=\frac{1}{i\omega C},\ \rho=-\frac{\omega^2LC}{4},$$

and

$$\cosh\,\Gamma=1-\tfrac{1}{2}\,\omega^2LC. \tag{226}$$

Now let us write $\Gamma=i\theta$, where $i=\sqrt{-1}$; the preceding equation becomes

$$\cos\,\theta=1-\tfrac{1}{2}\,\omega^2LC \tag{227}$$

and the *ratio of currents* in adjacent sections is $e^{-i\theta}$. *Consequently if θ is a real quantity the ratio of the absolute values of the currents in adjacent sections is unity, and the current is propagated without attenuation.*

Inspection of equation (227) shows that θ is real provided the right hand side lines between $+1$ and -1: or that ω lies between 0 and $2/\sqrt{LC}$. Consequently this type of artificial line transmits, in the steady state, sinusoidal currents of all frequencies from zero to $1/\pi\sqrt{LC}$ without attenuation. It is known as the low-pass filter.

If we invert the structure, that is, make the series impedance z_1 a capacity C and the shunt impedance z_2 an inductance L, so that

$$z_1=\frac{1}{i\omega C},\ z_2=i\omega L,\ \rho=-\frac{1}{4\omega^2LC},$$

we get, corresponding to (226) and (227),

$$\cosh\Gamma=1-\frac{1}{2\omega^2LC}, \tag{228}$$

$$\cos\theta=1-\frac{1}{2\omega^2LC}. \tag{228a}$$

This type of artificial line transmits without attenuation currents of all frequencies for which the right hand side of (228-a) lies between $+1$ and -1; that is, all frequencies from infinity to a lower limiting frequency $1/4\pi\sqrt{LC}$, while it attenuates all frequencies below this range. It is known, on this account, as the high-pass filter.

It is possible by using more complicated impedances to design filters which transmit a series of bands of frequencies. We cannot, however, go into the complicated theory of wave filters here, which has been covered in a series of important papers. One point should be noted, however: transmission without attenuation implies that the impedance elements are non-dissipative. Actually, of course, all the elements introduce some loss, so that in practice the filter attenuates all frequencies. Careful design, however, keeps the attenuation very low in the transmission bands.

We shall now derive the indicial admittance formulas for some representative types of artificial lines and wave filters from the operational formula

$$A_n=\frac{1}{\sqrt{(1+\rho)z_1z_2}}[\sqrt{1+\rho}+\sqrt{\rho}\,]^{-2n}. \tag{229}$$

This equation follows directly from (225) on putting $V_o=1$.

We start with the so-called low-pass filter on account of its simplicity and also its great importance in technical applications. This type of filter consists of series inductance L and shunt capacity C. The general case which includes series resistance R and shunt leakage G has been worked out (see Transient Oscillations, Trans. A. I. E. E., 1919). The solution is, however, extremely complicated and will not be dealt with here. We shall, instead, consider the important and illuminating case where the series and shunt losses are so related

as to make the circuit quasi-distortionless. We therefore take, operationally,

$$\begin{aligned} z_1 &= pL+R = L(p+\lambda) \\ 1/z_2 &= pC+G = C(p+\lambda) \end{aligned} \tag{230}$$

where $\lambda = R/L = G/C$.

We then have

$$\begin{aligned} z_1 z_2 &= L/C, \\ z_1/z_2 &= LC(p+\lambda)^2, \\ \rho &= \frac{LC}{4}(p+\lambda)^2. \end{aligned} \tag{231}$$

Now by reference to formula (229) we see that A_n is a function of $(p+\lambda)$; thus

$$A_n = \frac{1}{Z_n(p+\lambda)} = \left(1+\frac{\lambda}{\rho}\right)\frac{\rho}{(\rho+\lambda)Z_n(\rho+\lambda)}.$$

Now write

$$A_n^o = \frac{1}{Z_n(p)}.$$

It follows at once from reference to theorem VII that

$$A_n = \left(1+\lambda\int_0^t dt\right)A_n^o e^{-\lambda t} \tag{232}$$

so that the problem is reduced to the solution of the operational equation for A_n^o. Writing $\omega_c = 2/\sqrt{LC}$, we have

$$\begin{aligned} A_n^o &= \sqrt{\frac{C}{L}}\,\frac{1}{\sqrt{1+(p/\omega_c)^2}}\left[\sqrt{1+(p/\omega_c)^2}+p/\omega_c\right]^{-2n} \\ &= \sqrt{\frac{C}{L}}\,\frac{\omega_c}{\sqrt{p^2+\omega_c^2}}\left[\frac{\sqrt{p^2+\omega_c^2}-p}{\omega_c}\right]^{2n}. \end{aligned} \tag{233}$$

Now refer to formula (n) of the table of integrals; writing $\sqrt{L/C}=k$, we see by Theorem V that

$$A_n^o = \frac{1}{k}\int_0^{\omega_c t} J_{2n}(\tau)\,d\tau \tag{234}$$

where $J_{2n}(\tau)$ is the Bessel function of order $2n$ and argument τ. We note also that this is the indicial admittance of the non-dissipative low-pass wave filter; that is, the current in the n^{th} section in response

to a unit e.m.f. applied to the initial section ($n=0$). From (232) and (234) it follows at once that

$$A_n = e^{-\lambda t}\frac{1}{k}\int_0^{\omega_c t} J_{2n}(\tau)d\tau + \frac{\lambda}{k}\int_0^{t} d\tau e^{-\lambda\tau}\int_0^{\omega_c\tau} J_{2n}(\tau_1)d\tau_1.$$

Integrating the second member by parts and noting that $A_n^o(o)=0$, this reduces to

$$A_n = \frac{1}{k}\int_0^{\omega_c t} e^{-\frac{\lambda}{\omega_c}\tau} J_{2n}(\tau)d\tau \qquad (235)$$

which is the indicial admittance formula for the quasi-distortionless low-pass filter, or artificial line.

Before discussing these formulas, it is of interest to derive the formula for A_n^o by power series expansion. Formula (233) can be written

$$A_n^o = \frac{1}{k}\left(\frac{\omega_c}{p}\right)^{2n+1}\frac{1}{\sqrt{1+(\omega_c/p)^2}\,[1+\sqrt{1+(\omega_c/p)^2}]^{2n}}.$$

This can be expanded in a series in inverse powers of p; thus

$$A_n^o = \frac{1}{k2^{2n}}\left\{\left(\frac{\omega_c}{p}\right)^{2n+1} - \frac{2n+2}{2^2 1!}\left(\frac{\omega_c}{p}\right)^{2n+3} + \frac{(2n+3)\,(2n+4)}{2^4 2!}\left(\frac{\omega_c}{p}\right)^{2n+5} - \cdots\right\}.$$

Replacing $1/p^n$ by $t^n/n!$ in accordance with the Heaviside Rule we get

$$A_n^o = \frac{2}{k}\left\{\frac{1}{(2n+1)!}\left(\frac{\omega_c t}{2}\right)^{2n+1} - \frac{2n+2}{1!(2n+3)!}\left(\frac{\omega_c t}{2}\right)^{2n+3} + \frac{(2n+3)\,(2n+4)}{2!(2n+5)!}\left(\frac{\omega_c t}{2}\right)^{2n+5} - \cdots\right\}. \qquad (235\text{-a})$$

This can be recognized as the power series expansion of (234).

The *artificial cable* is also of interest and practical importance. In this structure the series impedance is a resistance R and the shunt impedance is a capacity C, so that

$$\begin{aligned} &z_1 = R,\ 1/z_2 = pC, \\ &z_1 z_2 = R/pC,\quad z_1/z_2 = pRC, \\ &\rho = pRC/4. \end{aligned} \qquad (236)$$

Now let us return to formula (229), and expand in inverse powers of ρ: we get

$$A_n = \frac{1}{2^{2n}\sqrt{\rho z_1 z_2}}\left\{\frac{1}{\rho^n} - \frac{2n+2}{2^2 1!}\frac{1}{\rho^{n+1}} + \frac{(2n+3)(2n+4)}{2^4 2!}\frac{1}{\rho^{n+2}} - \cdots\right\} \quad (237)$$

Now since $\sqrt{\rho z_1 z_2} = \frac{R}{2}$, we have

$$A_n = \frac{2}{2^n R}\left\{\left(\frac{2}{RCp}\right)^n - \frac{2n+2}{2.1!}\left(\frac{2}{RCp}\right)^{n+1} + \frac{(2n+3)(2n+4)}{2^2 2!}\left(\frac{2}{RCp}\right)^{n+2} - \cdots\right\}.$$

Replacing $1/p^n$ by $t^n/n!$ we get finally

$$A_n = \frac{2}{2^n R}\left\{\frac{1}{n!}\left(\frac{2t}{RC}\right)^n - \frac{(2n+2)}{2.1!(n+1)!}\left(\frac{2t}{RC}\right)^{n+1} + \frac{(2n+3)(2n+4)}{2^2.2!(n+2)!}\left(\frac{2t}{RC}\right)^{n+2} - \cdots\right\}. \quad (238)$$

For large values of n and t this series is difficult to compute or interpret. It can, however, be recognized as the series expansion of the function

$$A_n = \frac{2}{R}e^{-\frac{2t}{RC}} I_n\left(\frac{2t}{RC}\right) \quad (239)$$

where $I_n(2t/RC)$ is the Bessel function I_n of order n and argument $(2t/RC)$. This solution, it may be remarked, can be derived directly by a modification of the integral formula (n).

It is beyond the scope of this paper to consider other types of artificial lines and wave filters; for a fairly extensive discussion the reader is referred to "Transient Oscillations in Electric Wave-Filters," B. S. T. J., July, 1923. The low-pass wave filter, however, both in its own right and on account of its close relation to the periodically loaded line, deserves further discussion.

For the non-dissipative low-pass wave filter, we have

$$A_n^o = \frac{1}{k}\int_0^{\omega_c t} J_{2n}(\tau)\,d\tau \quad (234)$$

while for the quasi-distortionless low-pass wave filter

$$A_n = \frac{1}{k}\int_0^{\omega_c t} e^{-\mu\tau} J_{2n}(\tau)\,d\tau \quad (235)$$

where $\mu = \lambda/\omega_c = R/L\omega_c = R/2vL$.

Computation and analysis of these formulas involve an elementary knowledge of Bessel functions. The properties necessary for our purposes are briefly discussed in an appendix to this chapter.

The indicial admittances for the non-dissipative low-pass filter,

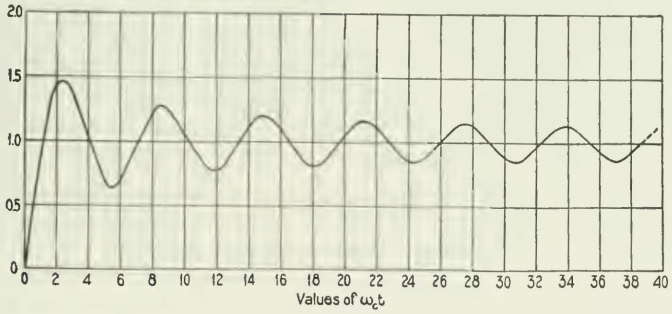


Fig. 20—Low pass wave filter. Indicial admittance of initial section ($n=0$). Multiply ordinates by $\sqrt{C/L}$

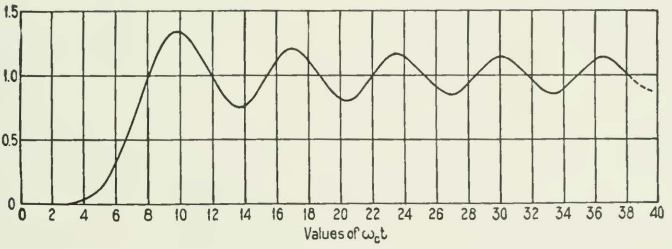


Fig. 21—Low pass wave filter. Indicial admittance of third section ($n=2$). Multiply ordinates by $\sqrt{C/L}$

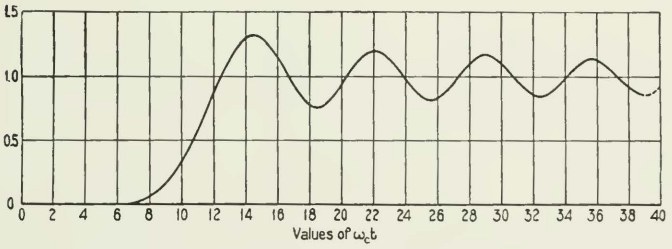


Fig. 22—Low pass wave filter. Indicial admittance of fifth section ($n=4$). Multiply ordinates by $\sqrt{C/L}$

that is, the current in response to a steady unit e.m.f. applied at time $t=0$, are shown in the curves of Figs. 20, 21 and 22, for the initial or zeroth, the 3rd and the 5th sections, respectively. These curves together with the exact and approximate formulas given

above are sufficient to give a reasonably comprehensive idea of the general character of these oscillations and their dependence on the number of sections and the constants of the filter.

It will be observed that the current is small until a time approximately equal to $2n/\omega_c = n\sqrt{L_1C_2}$ has elapsed after the voltage is applied. Consequently the low-pass filter behaves as though currents were transmitted with a finite velocity of propagation $\omega_c/2 = 1/\sqrt{L_1C_2}$ sections per second. This velocity is, however, only apparent or virtual since in every section the currents are actually finite for all values of time >0.

After time $t = n\sqrt{L_1C_2}$ has elapsed the current oscillates about the value $1/k$ with increasing frequency and diminishing amplitude. The amplitude of these oscillations is approximately

$$\frac{1/k}{\sqrt{1-(2n/\omega_c t)^2}}\sqrt{\frac{2}{\pi\omega_c t}}$$

and their instantaneous frequency (measured by intervals between zeros)

$$\frac{\omega_c}{2\pi}\sqrt{1-(2n/\omega_c t)^2}.$$

The oscillations are therefore ultimately of cut-off or critical frequency $\omega_c/2\pi$ in all sections, but this frequency is approached more and more slowly as the number of filter sections is increased.

Figs. 23, 24, 25, give the indicial admittance in the 100th, 500th and 1000th section of the low-pass wave filter. The filter itself seldom

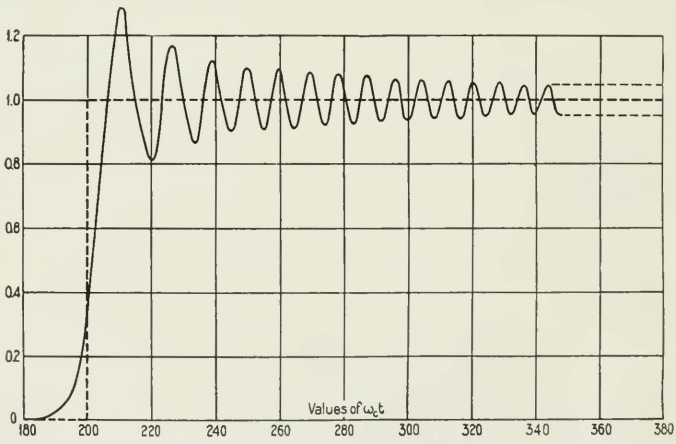


Fig. 23—Low pass wave filter. Indicial admittance of 100th section ($n=99$). Multiply ordinates by $\sqrt{C/L}$

embodies more than 5 sections. The case of a large number of sections is of interest, however, because it represents a first approximation to the periodically loaded line. While the non-dissipative

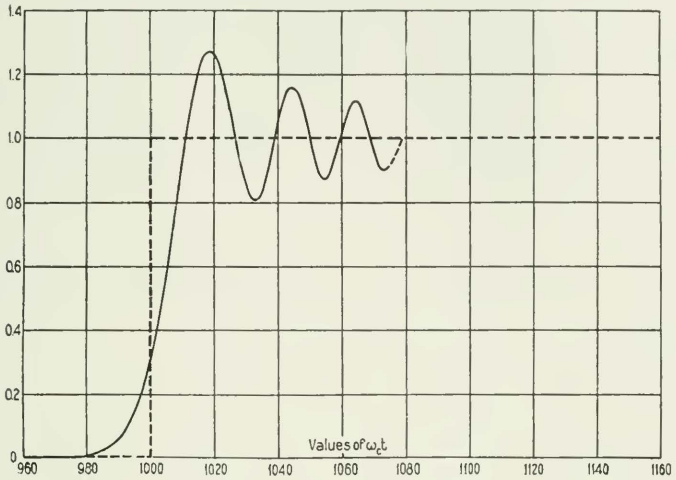


Fig. 24—Low pass wave filter. Indicial admittance of 500th section ($n=499$). Multiply ordinates by $\sqrt{C/L}$

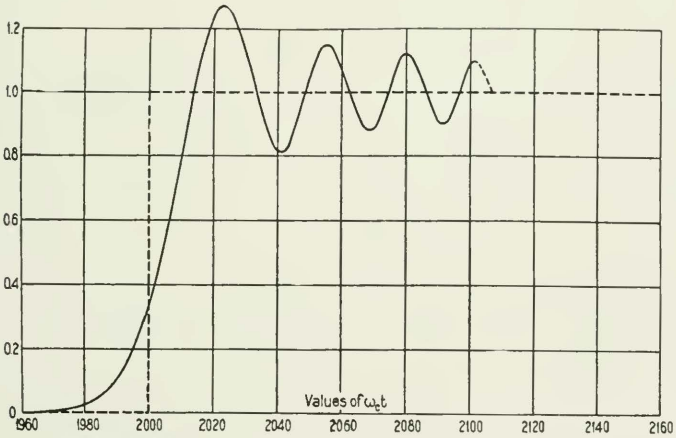


Fig. 25—Low pass wave filter. Indicial admittance of 1000th section ($n=999$). Multiply ordinates by $\sqrt{C/L}$

line is ideal and unrealizable, its study is of practical importance because in this type of line the effect of the discontinuous character of the loading of the periodically loaded line is isolated and exhibited in the clearest possible manner.

The dotted curves represent the current in the corresponding smooth line. For the smooth line, the current, as we have seen, is discontinuous, being identically zero for a time $vt=n$ and having an instantaneous jump to its final value $\sqrt{C/L}$ at $vt=n$. The current in the artificial or periodically loaded line differs from that in the corresponding smooth line in three important respects: (1) the absence of the abrupt discontinuous wave front, (2) the presence of superposed oscillations, and (3) *the absence of a true finite velocity of propagation.* It will be observed, however, that the current in any section is negligibly small or even sensibly zero until $vt=n$, so that the current is propagated with a *virtual* velocity $1/\sqrt{LC}$ per section. The presence of a well marked wave front is also evident although this is not abrupt, as in the smooth line. The effective slope of the wave front becomes smaller as the current wave travels out on the line, decreasing noticeably as the number of sections is increased. When the number of sections becomes large, however, the decrease in the slope is not rapid, being in the 500^{th} section about 60 per cent. of that in the 100^{th} section.

The superposed oscillations are of interest. These are initially of a frequency depending upon and decreasing with the number of sections, n, but in all sections ultimately attaining the frequency

$$\frac{1}{\pi\sqrt{LC}}=\frac{v}{\pi}$$

which is the critical or cut-off frequency of the line, above which steady-state currents are attenuated during transmission and below which they are unattenuated. When vt is large compared with n the amplitude of these oscillations becomes $\sqrt{1/\pi vt}$ so that they ultimately die away and the current approaches the value $\sqrt{C/L}$ for all sections. The current in the loaded line is thus asymptotic to the current in the corresponding smooth line and oscillates about it with diminishing amplitude and increasing frequency.

Since the abscissas of these curves represent values of $2vt=2t/\sqrt{LC}$, and the ordinates are to be multiplied by $\sqrt{C/L}$ to translate into actual values, the curves are of universal application for all values of the constants L and C.

The investigation of the building-up of alternating currents in wave filters and loaded lines is very important. It depends for the non-dissipative case on the properties of the definite integrals

$$\int_0^{\omega_c t} \sin w\tau\, J_n(\tau)d\tau,$$

$$\int_0^{\omega_c t} \cos w\tau\, J_n(\tau)d\tau,$$

where $w=\omega/\omega_c$ and $\omega=2\pi$ times the applied frequency. The mathematical discussion is, however, quite complicated and will not be entered into here. The reader, who wishes to follow this further, is referred to Transient Oscillations, Trans. A. I. E. E., 1919 and Transient Oscillations in Electric Wave Filters, B. S. T. J., July, 1923.

Appendix to Chapter VIII. Note on Besel Functions

The Bessel Functions of the first kind, $J_n(x)$ and $I_n(x)$, are defined, when n is zero or a positive integer, by the absolutely convergent series

$$J_n(x)=\frac{x^n}{2^n.n!}\left\{1-\frac{x^2}{2(2n+2)}+\frac{x^4}{2.4(2n+2)(2n+4)}-\frac{x^6}{2.4.6.(2n+2)(2n+4)(2n+6)}+\cdots\right\},$$

$$I_n(x)=\frac{x^n}{2^n.n!}\left\{1+\frac{x^2}{2(2n+2)}+\frac{x^4}{2.4(2n+2)(2n+4)}+\frac{x^6}{2.4.6.(2n+2)(2n+4)(2n+6)}+\cdots\right\}.$$

In the following discussion of the properties of these functions it will be assumed that the argument x is a pure real quantity.

For large values of the argument (x large compared with n), the behavior of the functions is shown by the asymptotic expansions:—

$$I_n(x)=\frac{e^x}{\sqrt{2\pi x}}\left\{1-\frac{4n^2-1}{1!\,(8x)}+\frac{(4n^2-1)(4n^2-9)}{2!\,(8x)^2}-\frac{(4n^2-1)(4n^2-9)(4n^2-25)}{3!\,(8x)^3}+\cdots\right\},$$

$$J_n(x)=\sqrt{\frac{2}{\pi x}}\left\{P_n\cos\left(x-\frac{2n+1}{4}\pi\right)-Q_n\sin\left(x-\frac{2n+1}{4}\pi\right)\right\},$$

where

$$P_n=1-\frac{(4n^2-1)(4n^2-9)}{2!\,(8x)^2}+\frac{(4n^2-1)(4n^2-9)(4n^2-25)(4n^2-49)}{4!\,(8x)^4}-\cdots,$$

$$Q_n=\frac{4n^2-1}{8x}-\frac{(4n^2-1)(4n^2-9)(4n^2-25)}{3!(8x)^3}+\cdots.$$

We thus see that I_n increases indefinitely and behaves ultimately as

$$\frac{e^x}{\sqrt{2\pi x}}.$$

The function $J_n(x)$, however, is oscillatory and ultimately behaves as

$$\sqrt{\frac{2}{\pi x}}\cos\left(x-\frac{2n+1}{4}\pi\right).$$

For all orders of n

$$\int_0^\infty J_n(x)dx=1.$$

The properties of $J_n(x)$ may be described qualitatively as follows:—

When the argument is less than the order ($0\leqq x<n$) the function is very small and positive, and is initially zero (except when $n=0$). In the neighborhood of $x=n$, the function begins to build up and reaches a maximum a little beyond the point $x=n$. Thereafter the function oscillates with increasing frequency and diminishing amplitude, and ultimately behaves as

$$\sqrt{\frac{2}{\pi x}}\cos\left(x-\frac{2n+1}{4}\pi\right).$$

When $n=0$, the initial value is unity, but the subsequent behavior of the function is as described above.

A more precise description of the function is gotten from the following approximate formulas.

$$J_n(x)\doteq B_n(x)\cos\Omega_n(x), \qquad \text{for } x>n$$

where

$$B_n(x)=\sqrt{\frac{2}{\pi x}}\frac{1}{\left(1-\frac{m^2}{x^2}+\frac{3}{2}\frac{m^2}{x^4}\frac{1}{(1-m^2/x^2)^2}\right)^{1/4}},$$

$$\Omega_n(x)=x\left[\sqrt{1-\frac{m^2}{x^2}}+\frac{m}{x}\sin^{-1}\left(\frac{m}{x}\right)-\frac{m^2}{4x^4}\frac{1}{(1-m^2/x^2)^{3/2}}\right]-\frac{2n+1}{4}\pi,$$

$$\Omega_n'(x)=\frac{d}{dx}\Omega_n(x),$$

$$=\sqrt{1-\frac{m^2}{x^2}+\frac{3}{2}\frac{m^2}{x^4}\frac{1}{(1-m^2/x^2)^2}},$$

and

$$m^2=n^2-1/4.$$

This approximate formula is valid only where $x>n$, its accuracy increasing with x and with n. For all orders of n it is quite accurate beyond the first zero of the function.

The "instantaneous frequency" of oscillation is approximately

$$\frac{1}{2\pi}\Omega_n'(x) = \frac{1}{2\pi}\sqrt{1-\frac{m^2}{x^2}+\frac{3}{2}\frac{m^2}{x^4}\frac{1}{(1-m^2/x^2)^2}}.$$

By this it is meant that at any point $x(x>n)$ the interval between successive zeros is approximately $\pi/\Omega'(x)$. Otherwise stated, in the neighborhood of any point x, the function behaves like a sinusoid of amplitude $B_n(x)$ and frequency $\omega/2\pi$ where $\omega=\Omega_n'(x)$.

The following approximate formulas, while not sufficiently precise for the purposes of accurate computation except for quite large values of x, clearly exhibit the character of the functions for values of the argument $x>n$, and of the order $n>2$.

$$J_n(x) \doteq h_n\sqrt{\frac{2}{\pi x}}\cos\ (q_n x-\theta_n),$$

$$J_n'(x) = -q_n h_n\sqrt{\frac{2}{\pi x}}\sin\ (q_n x-\theta_n),$$

$$\int_0^x J_n(x)dx = 1+\frac{h_n}{q}\sqrt{\frac{2}{\pi x}}\sin\ (q_n x-\theta_n),$$

where

$$h_n=\left(\frac{1}{1-n^2/x^2}\right)^{1/4}=1+\frac{n^2}{4x^2},$$

$$q_n=\sqrt{1-n^2/x^2},$$

$$\theta_n=\frac{2n+1}{4}\pi-n\sin^{-1}(n/x).$$

Some Contemporary Advances in Physics—X
The Atom-Model, Third Part

By KARL K. DARROW

M. A Very Brief Recapitulation of What Has Gone Before

ABUNDANT evidence of many kinds exists to show that each and every distinct sort of atom is especially adapted to possess energy, not in any random quantity whatsoever, but in certain peculiar, definite, characteristic amounts. An atom having energy in one of these particular amounts apparently cannot add arbitrary quantities to its store, nor yield up arbitrary quantities from it; whenever the atom receives or whenever it gives energy, it receives or gives only just so much as is exactly sufficient to raise or reduce its supply to some one among the others of these distinctive quotas. For each of the chemical elements there is a great system of these distinctive energy-values. They are determined chiefly by analyzing spectra, and for most of the elements—the exceptions being those of which the spectra are excessively complicated—many of them have been evaluated very accurately and set down in tables. The system of distinctive energy-values for any element is a very important feature of that element; perhaps, indeed, the most important feature of all.

It is customary to say that when an atom acquires or surrenders energy, it passes from one into another state; the various states corresponding to its various distinctive energy-values are called its "Stationary States." This is a name which suggests, and is doubtless meant to suggest, that the energy-value of the atom is but one among many of its features, all of which change when the energy-value changes. This is a legitimate idea; theorizing about the atom consists in speculating about just such features. But the reader will go far and grievously astray if he lets the name signify to him that many of them are directly and definitely known. In some few cases there is good reason to believe that we know the magnetic moment of an atom in its normal state. Beyond these the energy-values are all that are known. If the reader chooses everywhere to replace "Stationary State" by "energy-value" he will be holding fast to the physical reality, to the one thing not liable to be compromised by the future trends of thought.

An atom may pass from one Stationary State to another because of colliding with an electron or another atom of the same or a different

kind; or by absorbing radiation it may pass from one Stationary State to another of higher energy; or it may pass spontaneously from one Stationary State to another of lower energy. In this last case, it emits radiation of which the frequency ν is related to the difference ΔU between the energy-values of the initial Stationary State and the final one by the equation

$$\nu = \Delta U/h, \qquad (h = 6.56.10^{-27} \text{ erg/sec.})$$

The same equation governs the last case but one, in which it connects the frequency of the absorbed radiation with the energy-difference between the two Stationary States from and into which the atom passes. On this equation is founded the method of analyzing spectra which is the most accurate and most widely applicable method of determining the energy-values of Stationary States. The other ways in which atoms are caused to pass from one State to another lead to methods of determining these states, which are almost useless for accurate measurements, but invaluable as controls.

The energy-values of the various Stationary States of an atom are interrelated, and sometimes it is possible to express a long sequence of them by means of a simple or a not very complicated formula. There are also interrelations between the distinctive energy-values for different elements; and this statement is meant to apply also to atoms from which electrons, one or more, have been detached, which should be considered as distinct though not as stable elements. There are unmistakable numerical relations among the Stationary States which come into being when atoms are subjected to electric or to magnetic fields. Finally there is the important principle that the spontaneous transitions between various pairs of states, which result in spectrum lines, do not occur equally often; and yet the relative oftenness or seldomness of their occurrence is itself regulated by laws. One finds instances in which transitions from a state A to a state B_1 are just twice as common as transitions from A to a state B_2 close to B_1. One finds instances in which transitions from a state A to a state B do not occur at all under usual conditions and an atom in state A cannot get into state B without touching at some state C from which A and B are both accessible. It is possible so to arrange the Stationary States of an atom that by looking at the situations of any two states in the arrangement, one can tell immediately whether direct transitions between them do occur or do not; and this arrangement is found to be suited to, even to be demanded by the numerical interrelations to which I alluded above. Upon these facts the classi-

fications of the Stationary States are founded, and the notations by which they are named.

The atom-model to which this article is devoted, the atom-model of Rutherford and Bohr, is designed to interpret these facts of the Stationary States, but not these alone. It is designed also to interpret certain experiments—chiefly, though not altogether, experiments on the deflections suffered by minute flying charged particles when they pass through matter—which indicate that an atom consists of a positively-charged nucleus with a congeries of electrons around it. Specifically, the results of these experiments agree with the notion that the Nth element of the Periodic Table consists of a nucleus with positive charge Ne and N electrons surrounding it; and this is the simplest and most satisfying notion with which they do agree. Yet there is something paradoxical about this atom-model; for electrons could neither stand still nor yet revolve permanently in orbits around a nucleus, if they conformed to the laws of electrostatics. Also there must needs be something paradoxical about any attempt to interpret the Stationary States by this model, for there is nothing inherent in it to make any energy-value preferable to any other. Under these circumstances Bohr's procedure was, resolutely to accept both paradoxes at once, and to say that the electrons can revolve permanently in those and just those particular orbits, whereby the energy of the atom assumes the particular values which are those of the observed Stationary States. This is easy to say; but it is not important, unless one succeeds in showing that those and only those particular orbits are set apart from all others by some peculiar feature, are distinguished by conforming to some particular principle, which can be exalted into a "Law of Nature" to complement or supersede the laws of electrostatics. Otherwise the atom-model would be of no value.

Thus in order to make the test of the atom-model, it is necessary to trace these orbits. One is confronted with this problem of orbit-tracing: *Given:* the observed energy-values of the Stationary States; *required:* to trace the orbits such that, when the electrons travel in them, the energy of the atom has these observed values. If this problem cannot be solved, it is impossible to take the next and essential step of ascertaining whether these particular orbits are distinguished in any particular way from all the other conceivable ones.

In the case of a single electron revolving about a nucleus, this problem is sometimes soluble. If the mass of the electron is regarded as invariable, and no outside influences are supposed to act upon the atom, then the solution is comparatively easy to attain. It was performed in the Second Part of this article. If an external magnetic

field is superposed, the problem is scarcely more difficult; if an external electric field is superposed, it is difficult but soluble—provided always that the mass of the electron be supposed invariable. If the mass of the electron is supposed to vary with its speed as the theory of relativity requires, and as certain experiments suggest, the problem remains soluble—provided that no outside influences act. For all these cases the orbits which yield the observed energy-values have been traced; and certain features have been shown to be common to all of these "permitted" orbits, and to no others, so that by these features the permitted orbits are set apart from all the rest. Inversely, anyone who is told what these features are, and who is sufficiently adept in dynamics, can trace all the orbits which display them and calculate the energy-values for these orbits and so predict the energy-values of all of the Stationary States of an atom consisting of a nucleus and a single electron. Such orbits are known as *quantized orbits*. The rules whereby they are set apart from all the multitude of orbits not permitted are the *Quantum Conditions*; which some one, it is to be hoped, will some day succeed in deriving from a general Principle of Quantization.

The most general way of phrasing these conditions is difficult to grasp, and the more intelligible ways are not the most general. The most general conditions yet formulated are not adequate for all cases; the completely adequate principle is yet to be discovered. For the purposes of this summary and of most of what follows, a very limited expression of the Quantum Conditions will be sufficient. In the Second Part of this article it was proved that the permitted orbits of an electron of invariable mass revolving in an inverse-square field such as is supposed to surround a bare nucleus, are certain ellipses. It was further stated, without proof, that if the electron of invariable mass revolves in a central field which deviates slightly from an inverse-square field, then the permitted orbits are certain "rosettes" or precessing ellipses—each orbit may be traced by imagining an ellipse revolving steadily in its own plane around the source of the central field (I will say the nucleus) at one of its foci. All the orbits are rosettes; the permitted orbits are certain rosettes which are distinguished from the others by a distinctive feature. One way of expressing this feature involves the angular momentum p_ϕ and the radial momentum p_r of the electron. In terms of the mass m of the electron, its distance r from the nucleus, and its angular velocity $d\phi/dt$ about the nucleus, these quantities are by definition:

$$p_r = m(dr/dt), \qquad p_\phi = mr^2(d\phi/dt). \tag{1}$$

The "principle of quantization" is, that the permitted orbits are marked out from all others in that they fulfil these conditions, which are the Quantum Conditions:

$$\int p_r \, dr = kh, \tag{2}$$

$$\int p_r \, dr + \int p_\phi \, d\phi = nh. \tag{3}$$

In these equations each integral is taken around one complete cycle of the corresponding variable; h stands for Planck's constant, and n and k take the values of all positive integers, k never surpassing n.

There is an alternative way of phrasing these quantum-conditions, which is much easier to visualize; but it emphasizes what are probably accidental features of the permitted rosettes, rather than fundamental ones. The rosettes are, as I have said, precessing ellipses; the major axes $2a$ and the minor axes $2b$ of these ellipses are, for the permitted rosettes

$$2a = n^2 h^2 / 2\pi^2 e^2 m, \tag{4}$$

$$2b = (k/n)\, 2a = k\, n\, h^2 / 2\pi^2 e^2 m, \tag{5}$$

in which n and k take as before the values of all positive integers, k not surpassing n.

Exactly the same principle governs the permitted orbits of an electron revolving in a perfect inverse-square central field, but varying in mass when its speed varies, as the theory of relativity requires. In this case also the orbits are rosettes, and the permitted orbits are particular rosettes set apart from all the others in that they fulfil (2) and (3), therefore automatically (4) and (5). The energy-values of these permitted orbits agree closely with those of the observed Stationary States of hydrogen and of ionized helium, the atoms of which are the only atoms believed to consist of a nucleus and one electron. Inversely, the orbits required to interpret the observed Stationary States are set apart from all the other conceivable orbits by the features expressed by (2) and (3), and by (4) and (5). On these close numerical agreements for hydrogen and ionized helium, and on other numerical agreements for the same atoms arising when external fields are applied, the prestige of Bohr's atom-model is founded.

The integers n and k, the *total quantum number* and the *azimuthal quantum number*, are used as indices to symbolize the various Stationary States of hydrogen and ionized helium to which they correspond. Thus the symbol "3_2" stands for the Stationary State of either atom,

which in the atom-model is realized when the electron circulates in a rosette, or precessing ellipse, for which $n=3$ and $k=2$. Orbits for which $k=n$ are circles; orbits for which $k<n$ are (precessing) ellipses, and the farther k falls below n the more eccentric (narrower) is the ellipse, although its major axis is independent of k. I have repeated all of these statements about precessing ellipses and their quantum-numbers, because a great part of the speculation about atoms possessed of more than one electron consists of persevering and obstinate attempts to interpret their behavior by as nearly as possible the same ideas.

It is essential to remember also that all the energy-values of the Stationary States are reckoned from the "State of the Ionized Atom," in which State the energy—i.e., the energy of a system composed of one atom deprived of an electron, and one electron far away—is equated to zero.

N. Introduction to the Speculations About Atoms with More than One Electron

All atoms, except those of hydrogen and ionized helium, possess more than one electron. There is much evidence of various kinds for this assertion; and certainly the spectra of these other atoms cannot be interpreted as those of the first two have been. Thus we are confronted with the problem of a system composed of a nucleus and more than one electron. The similarities between the spectra of hydrogen and ionized helium, and those of other elements, are important enough to make it desirable to use the same sort of explanation. We imagine the various electrons, when there are two or more, each to describe certain permitted orbits, set apart from the multitude of other conceivable orbits by peculiar features expressible by a Principle of Quantization.

Here at the outset we meet with the great hindrance to success in this problem. It is not possible to determine what features are common to permitted orbits, for it is not possible even to trace the permitted orbits. The general problem of tracing the paths of three or more bodies, attracting or repelling one another according to the inverse-square law, remains unsolved. Considering that for centuries the related but simpler problem of celestial mechanics has been under continual and powerful attack, the general problem may fairly confidently be regarded as insoluble. There is very little hope of ever dominating it to such an extent, that the spectra of atoms with two or more electrons can be interpreted exactly by Bohr's atom-model,

or can be used as strong support to that theory. If those two spectra of hydrogen and ionized helium were unknown, it is unlikely that the atom-model would ever have been suggested; it is more than unlikely that the atom-model could ever have been regarded as satisfactory. To this day the prestige of the atom-model results almost entirely from its achievement with those two spectra.

Why then trouble with applying it to the interpretation of other spectra? Several good reasons can be given. For instance, it may be that a system of several electrons about a nucleus acts in some respects as a unit—that its motion can be considered in some ways as the motion of a rigid body, that principles of quantization can be found for the system as a whole, similar to the principles used for quantizing the smaller and yet perhaps not more consolidated system which a single electron is. Here and there in the discussion we shall find indications that this way of thinking is suitable.

Again, one is justified in arguing that if in simple cases a certain law is proved, and if in complex cases neither that nor any other law can be proved nor disproved, than we should assume that the law proved for the simple cases extends over the complex ones. Few events in this world take place under such conditions that conservation of energy can be proved to prevail during them; yet, from the fact that conservation of energy has been verified in whatever events it has been tested, we do not hesitate to infer that it prevails in all. Bohr's model having been so strongly fortified by the data for the only two atoms for which it can be completely tested, why not assume it for the others?

And finally, there is the point that many of the data of experiment are almost universally expressed in terms of the model, so that the physical literature of today is almost incomprehensible without some knowledge of it. Unfortunate as this is, it shows that the model is a valuable aid for visualizing the facts. This justifies any model; but must not be construed as evidence for it.

It will be expedient to divide the subject substantially under these following headings.

(a) *The Helium Atom.* This, as the case of an atom composed presumably of a nucleus and two electrons, comes nearest to being amenable to calculation. Certain mechanically possible orbits of the two electrons, possessing the peculiar features of the "permitted" orbits of a single electron revolving around a nucleus, have been traced and their energy-values calculated. Not one of them has given the observed energy-value of a Stationary State of the helium atom. It is the consensus of opinion that whatever the features

which distinguish the permitted orbits may be, they are not those which prevail in the hydrogen atom.

(b) *Alkali-Metal Atoms.* For these there is reason to believe that one electron is normally located far beyond all the others, and may be supposed to revolve around a "residue" consisting of all the others and the nucleus. At a great distance, the field due to this residue will be very nearly a central field such as would surround a nucleus of charge $+e$,—a hydrogen nucleus; for, from a great distance, the nucleus and the electrons of the residue will seem almost to coincide in place. Nearer in, the forces due to the electrons of the residue may be supposed to compound with that due to the nucleus in such a way that a central field, not varying as the inverse square, results. Thus rosette orbits may be expected (for this reason I quoted the principle of quantization for such orbits in Section M). An enormous amount of effort has been spent in constructing central fields, such that the rosette orbits obeying the quantum-conditions (2) and (3) have nearly the energy-values which the Stationary States of these atoms are known to possess. Always, the emission of a spectrum line is supposed to result from a transition of the outermost or valence electron from one orbit to another, the electrons of the residue being scarcely or not at all affected. Such is the general explanation for the far-reaching and yet imperfect resemblance of the spectra of these metals to that of hydrogen.

(c) *Other Elements.* As one passes across the Periodic Table from left to right along any row, the spectra rapidly lose resemblance to the hydrogen spectrum. This is taken to mean that the assumption used for alkali metals—the assumption that one electron lies far beyond the others, and executes transitions while the others remain unaffected—departs progressively further from the truth. Evidence exists that simultaneous transitions of two electrons occur, and very likely yet more drastic rearrangements taking place *en bloc.*

(d) *Building of Atoms by Consecutive "Binding" of Electrons.* An atom composed, when complete, of Z electrons arranged about a nucleus bearing the charge $+Ze$, may have been formed originally in Z stages by the consecutive advent of Z electrons, the first annexing itself to the bare nucleus, the second joining itself to the system composed of the nucleus and the first, and so on until as many have arrived as the nucleus is able to hold. Each of these stages should be accompanied by the emission of lines belonging to a particular spectrum; the ordinary hydrogen spectrum accompanies the formation of a hydrogen atom by the step-by-step binding of an electron to a nucleus of charge e, the ionized-helium spectrum accompanies

the joining of the first electron to a nucleus of charge $+2e$, the neutral-helium spectrum the adhesion of the second electron to such a nucleus. Spectra corresponding to the latest four or five stages, in the formation of atoms having many electrons when completed, have been observed. To a certain extent, but not entirely, an atom with Z electrons and a nuclear charge Z resembles an atom with Z electrons and a nuclear charge $Z+1$. To a certain extent, therefore, each atom in the Periodic System may be regarded as resembling the last stage but one in the formation of the next following atom. This fact is important in the interpretations of the Periodic Table.

(e) *Multiplets.* We next take account of the fact that the sequences of Stationary States, mentioned in the elementary theory and description of spectra, are actually sequences of groups of Stationary States; and inquire what may be supposed to differentiate the several states of a group from one another. An elaborate formal theory is based on the assumption that all of the electrons of what I have called the "residue" of the atom revolve, if not literally as a rigid block, at least with a resultant angular momentum which itself is quantized; and that the outermost electron revolves in its own orbit around this residue, the different Stationary States of the group differing from one another in respect of the inclination of the orbit to the axis of rotation of the residue. The theory is not quite coherent with what has gone before; and for that reason the reader should try to separate its essential qualities from its accidental ones.

(f) *Magnetic Properties of Atoms.* A magnetic field should treat a system of electrons revolving around a nucleus in the same way as it treats one electron, as was said in the Second Part of this article. One would expect that in this case, if in any, the behavior of complex atoms would resemble that of the hydrogen atom; yet there is a striking and inexplicable contrast. This, like the spectrum of the helium atom, shows that either the quantum conditions governing the hydrogen atom are not universal, or the expressions hitherto found for the quantum conditions are too limited. From the responses of atoms to magnetic fields something is learned about the magnetic properties of atoms and their residues, some part of which can be tested by direct experiment; and these experiments include what are probably, all things considered, the most perplexing and fascinating ones of recent years.

(g) *Interpretation of X-ray Spectra.* X-ray spectra are analyzed as other spectra are, and each absorption and each emission of an X-ray by an atom is associated with a transition between two Stationary States; these "X-ray Stationary States" however are distin-

guished from the others, by the circumstance that every one of them involves the absence of an electron from the atom; consequently they may be described as Stationary States of an atom-residue. There is reason to believe that each distinct State involves the absence of a particular one, or of one out of a particular group, of the electrons bound to the nucleus during the earlier stages of the imagined building of the atom by successive "binding" of electrons. The speculations about X-ray spectra consist largely in attempts to correlate the individual States with absences of particular electrons.

O. The Helium Atom

The problem of the nucleus with two electrons, the "dilemma of the helium atom" as van Vleck calls it, is one of the most tantalizing in contemporary physics. One feels confident *a priori* that the same quantum conditions as suffice so beautifully to constrain the one-electron atom to yield the hydrogen spectrum should also suffice, when applied to the orbits of two electrons, to yield the spectrum of neutral helium. Yet the various pairs of orbits conforming to these quantum conditions, which have already been traced, have been shown (with vast expenditure of intellectual labor by some of the ablest mathematical physicists of our time) to entail energy-values for the Stationary States which are hopelessly incorrect.

For instance, one might assume that when the helium atom is in its Normal State, the two electrons are revolving in a common circular orbit about the nucleus, being at each instant located at opposite ends of a diameter; and that this permitted orbit is determined by the condition that the angular momentum of each electron, or perhaps that of both together, is $h/2\pi$. This seems an obvious generalization of the Quantum Conditions for hydrogen; but it yields a false energy-value for the Normal State; and there is nothing more to be said. Kemble and van Vleck demonstrated that no arrangement in which the two electrons are symmetrically placed relatively to a line through the nucleus entails the proper energy-value for the normal state. This still leaves open the possibility that the two electrons are unsymmetrically placed—a possibility which to some people seems repellent enough to be excluded. Born and Heisenberg calculated the energy-values corresponding to pairs of orbits, one of which lies far beyond the other at all points, and both of which are concordant with the Quantum Conditions. These ought to have agreed with the energy-values of the Stationary States which are remote from the normal state and near the state of the ionized

atom; but they did not. This result is commonly regarded as the strongest evidence for the belief that the Quantum Conditions valid for an atom with one electron are not valid for an atom with two.

The atom-model favored by Kramers, and hence presumptively by Bohr, to represent a helium atom in its normal state, involves two electrons moving in orbits which are not coplanar nor even plane. Planes tangent to the two surfaces upon which the orbits are traced intersect each other at 120° along a line passing through the nucleus, and the electrons pass simultaneously across this line at opposite crossing-points. These orbits conform to the Quantum Conditions; and the resultant of the angular momenta of the two electrons, which is the angular momentum of the entire atom, is equal to $h/2\pi$. This atom-model likewise fails to have the right energy-value for the normal state.

P. Interpretation of the Optical Spectra of Alkali-Metal Atoms

The alkali metals (lithium, sodium, potassium, rubidium and caesium) are elements of which the atoms are easily deprived of a single electron apiece; one electron of each atom is, as the phrase goes, exceptionally "loosely bound." Many facts combine to indicate this; for instance, each of these elements enters with violence into chemical combinations, and the compounds which each forms are such as to suggest that its atom yields up one electron to the atom or atoms which join with it. Again, when a salt of one of these metals is dissolved, the molecules split up and the atoms of the metal are left wandering around in the solvent minus one electron apiece, while the atoms of the other element each hold on to one captured electron. More definite yet is the direct evidence that the ionizing potentials of the alkali metals are lower than those of any other elements in the same rows of the Periodic Table, those of rubidium and caesium being altogether the smallest known. These alkali metals follow, in the Periodic Table, immediately after the five noble gases helium, neon, argon, xenon and krypton respectively. These gases are chemically all but absolutely inert, almost never entering into combinations. Their ionizing-potentials are higher than those of any other elements in their respective rows of the periodic table, and those of helium and neon are the greatest known. The atoms of each of the alkali metals are much larger than those of the preceding inert gas.

From all these facts the inference is drawn, that the atom of each inert gas consists of a nucleus and electrons, at least the outermost

ones of which are arranged in a peculiarly stable and symmetrical fashion (as for instance, in a group of eight at the corners of a cube, though this is by no means sure); while the atom of the next following alkali metal consists of just this sort of arrangement or "inert-gas shell," now to be known as the "residue" or "kernel," and of one additional electron now to be known as the "valence-electron," usually much farther away from the nucleus.

If to such an atom-model we apply the doctrine of Stationary States, we may infer that for each and every arrangement of the electrons in the residue, or (to use more general terms) for each condition of the residue, there is a whole system of Stationary States differing from one another only in that the valence-electron travels in different ones among a system of quantized orbits. These orbits we may suppose to conform to the quantum-conditions (2) and (3), at least until convincing evidence is brought to the contrary. Such in fact is the interpretation of the system of Stationary States, transitions between pairs of which are responsible for the "optical spectrum" of each alkali metal.

An electron at a very great distance from the kernel of such an atom will experience an attraction towards it, practically indistinguishable from the attraction which would be exerted by a single (hydrogen) nucleus of charge $+e$. One might say that the $(Z-1)$ electrons surrounding the nucleus of charge $+Ze$ effectively cancel a portion $+(Z-1)e$ of the nuclear charge; or to use a more common word, that they "screen" it. As the imagined distant electron moves inward towards the kernel, the screening will cease to be perfect. An effect should occur analogous to the "stray field" which penetrates the meshes of a grid; since the electrons of the kernel do not form a continuous shell of electricity enclosing the nucleus, the latter should make itself felt through the interstices, although this effect may be diminished by the swift motion of the electrons. All this is speculation of the wildest kind. The only deduction reasonably safe is this, that very far outside the kernel the field will be very nearly the inverse-square field due to a hydrogen nucleus of charge $+e$; very near to the kernel the field will be quite incalculable [1]; while in between the very-far-out and the very-near-in region, there will be an intermediate region, in which there may be some chance of finding an adequately approximate expression for the field. On the existence of such a region, in which such an approximation is good enough to be valuable,

[1] Unless it is violently simplified by some agency or restriction of which at present we know nothing.

rests all the present hope of achieving numerically valid theories in this division of atomic physics.

One agreement between this theory and certain data may be demonstrated without making any specific approximation. The farther away the valence-electron remains from the kernel, the more nearly identical with the field of a hydrogen nucleus is the field in which it revolves, the more nearly should it behave like the electron of a hydrogen atom. Consider for instance, in a hydrogen atom, the orbit which yields the Stationary State for which n the total quantum-number and k the azimuthal quantum-number are both equal to 5. This orbit is a circle of which the radius is 10^{-7} cm.; far larger than the radius of any inert-gas atom, presumably *a fortiori* far larger than the kernel of any alkali-metal atom. Were the valence-electron of such an atom to describe this circle, it would pass everywhere in a field very nearly like that of a hydrogen nucleus, and should very nearly conform to the quantum conditions for this field. It follows that an orbit drawn in the actual field, obeying the quantum-conditions $n=5$ and $k=5$, would be very nearly such a circle with very nearly the same energy-value. The inference is drawn that for high values of n and k, the Stationary States of an alkali-metal atom should be very nearly identical with those of hydrogen. These orbits which lie far out from the kernel of the alkali metal atom, or from the nucleus of the hydrogen atom, have small energy-values. It may therefore be said that if we tabulate the Stationary States of the two atoms in order of decreasing energy-value, then the farther along the two tabulations we go, the more nearly should the two systems of Stationary States coincide.

This is found to be true, under a limitation. The limitation is an important aid in interpreting the arrangement of the Stationary States. It will be recalled from the First Part of this article that the Stationary States of the sodium atom are arranged in several sequences (there illustrated as columns in Fig. 7) known as the *s*-sequence and the *p*-sequence and the *d*-sequence and the *f*-sequence and others; and to these sequences successive values 1,2,3,4 . . . of a symbol k were appended. One basis for this classification is that when it is made, the occurrence or non-occurrence of transitions between any pair of Stationary States, under normal conditions, can be determined by applying the "selection-rule" that only such transitions occur as involve a change of one unit in k. Now there are two reasons for supposing that the only transitions which can occur are those in which the Azimuthal Quantum-number of the valence-electron changes by one unit. Unfortunately it is not possible to introduce these two

reasons with all the necessary background without too long a stoppage of the main current of this argument.[2] I must therefore set it down as an assertion, that the selection-rule is deducible from the assumption that the value of k is the Azimuthal Quantum-number of the valence-electron; which thus is 1 for all the Stationary States of the *s*-sequence, 2 for each State belonging to the *p*-sequence, 3 for the *d*-sequence, and 4 for the *f*-sequence. The feature common to the various Stationary States of a sequence is, therefore, the Azimuthal Quantum-number of the valence-electron—if this atom-model is valid.

This being assured, the conclusion is drawn that, since k is higher for the *f*-terms than for the *d*-terms, higher for the *d*-terms than for the *p*-terms and higher for the *p*-terms than for the *s*-terms; since, therefore, the *f*-orbits are *ceteris paribus* more nearly circular than the *d*-orbits and less inclined to stretch down into the neighborhood of the kernel, the *d*-orbits more nearly circular than the *p*-orbits and the *p*-orbits more nearly circular than the *s*-orbits—therefore the approximation of the sodium terms to the hydrogen terms will be most nearly perfect for the *f* (and higher) sequences, less so for the *d*, less for the *p* and worst for the *s*-terms. This also is verified. It reinforces the opinion that the k-values assigned to the various sequences are actually their azimuthal quantum-numbers.

As the different States of a single sequence share a common Azimuthal Quantum-number, they must differ—supposing always that this atom-model is valid—in their Total Quantum-number. Consecutive States of a sequence presumably have consecutive values of the Total Quantum-number (although sometimes one meets with a break or a jolt in the continuity of a sequence, suggesting a departure from this rule). The meanings of the Total Quantum-number n and of the Azimuthal Quantum-number k for elliptical orbits are such, that n can never be less than k. Hence the value of n for the first Stationary State of the *s*-sequence may be unity, or greater; but the values of n for the first terms of the *p*-sequence, the *d*-sequence and the *f*-sequence may not be less than 2, 3, and 4, respectively.

Strange as it may seem, there is no perfectly satisfactory way of determining the value of n for all Stationary States. Generally it happens that the various States of an *f*-sequence, that of sodium for example, agree so closely with those States of hydrogen which form an n_4 sequence, that there is little hesitation in attaching to each of the *f*-States the same value of n as is borne by that State of the hydrogen atom which coincides with it so nearly. For instance,

[2] These being (*verbum sapienti*) the argument associated with the name of Rubinowicz, and the argument deduced from the Principle of Correspondence.

the first f-State of sodium has very nearly the same energy-value as the 4_4 State of hydrogen; the second f-State of sodium nearly coincides with the 5_4 State of hydrogen, and so forth along the sequence. Hence to the successive States of the f-sequence of the sodium atom one attaches with confidence the symbols 4_4, 5_4, 6_4, and so onward. In some cases this is practicable for the terms of the d-sequence also; but never for those of the s-sequence. The Stationary States of the s-sequence depart so far from those of hydrogen, that one cannot with any security conclude what values of the Total Quantum Number should be assigned to them. It used to be assumed that $n=1$ for the first term of the s-sequence and $n=2$ for the first term of the p-sequence, and the usual notation for the Stationary States reflects this supposition; which however is neither necessary nor probable.

All of the foregoing interpretations are based upon a theory of the alkali-metal atoms which may be summarized in this way: as the hydrogen atom is supposed to consist of a nucleus surrounded by an inverse-square field through which an electron travels always in one or another of certain orbits determined by quantum-conditions, so also the alkali-metal atom is supposed to consist of a kernel surrounded by a not-inverse-square field through which an electron travels always in one or another of certain orbits determined by identical quantum-conditions. As the Stationary States of the hydrogen atom correspond each to a certain orbit and are designated each by certain values of two quantities n and k, or for short by a symbol n_k indicating the features of that orbit, so also the Stationary States of the alkali-metal atom correspond each to a certain orbit and are designated each by a symbol n_k. For the hydrogen atom we recognize the proper n_k for each Stationary State because of the wonderful numerical agreement between Bohr's theory and the experimental values for the energy of each State. For the alkali-metal atom we can only guess the proper n_k for each Stationary States from indications of much lesser evidential value. We suppose, however, that $k=1,2,3,4$ for the various States of the s, p, d and f sequences, respectively; so that the s-sequence is like the n_1 sequence of hydrogen, the p-sequence like the n_2 sequence, and so on. Of the values of n we are moderately sure for the f and d sequences, quite uncertain for the terms of the s and p sequences.

One may now wonder whether it is possible to invent a central field, such that the orbits traced in it according to the quantum-conditions (2) and (3) would yield a series of energy-values agreeing with the observed energy-values of the Stationary States of (let me

say) the sodium atom. It takes a certain amount of faith to go about the business of designing such a central field; for the model imagined for the sodium atom involves ten electrons around the nucleus in addition to the one "valence" electron for the benefit of which the field is being devised; and one might expect these ten electrons to be rushing around the nucleus in uncoordinated and non-recurring paths, never at any two instants similarly placed and similarly moving, never at any two instants exerting the same influence upon the valence-electron. Yet the Stationary States of the sodium atom are as sharply defined as those of the hydrogen atom; and this may be thought to mean that the ten electrons of the kernel are constrained to a unity and a fixed relationship, like that of the members of a machine if not like that of the parts of a rigid body, which translates itself into an influence upon the valence-electron not unlike that of a central field.

At all events, several physicists working independently in various nations have taken the not inconsiderable trouble of devising central fields to fulfil the condition required; and they appear to have achieved a respectable success. It is not easy to decide what this success requires the rest of us to believe; perhaps it is formally possible to devise a central field to account for *any* set of Stationary States; I am not sure whether this question has been adequately examined. Some have felt confident enough to say that the results show which of the Stationary States correspond to orbits of the valence-electron which "penetrate into the kernel" and which to orbits that remain in all their circuit quite outside of the kernel. It is to be hoped that this problem will become clearer in the next few years. At this point I will add only, that the orbits traced for the valence-electron are rosette orbits in which the precession is very rapid, so that consecutive loops of a rosette are inclined at a considerable angle to one another. In the model for the hydrogen atom, the consecutive loops of a rosette orbit lie so close together as to be indistinguishable when drawn to scale on an ordinary sheet of paper (the separation between them was much exaggerated in Fig. 3 of the Second Part of this article). In these atom-models, the orbit looks rather as if it were drawn along the edges of the blades of an electric fan.

Q. Interpretation of the Optical Spectra of Other Elements

As soon as we step from the first column of the Periodic Table into the second, the obstacles to such a theory as we have hitherto tried to hold are gravely increased. There is evidence of several kinds

which seems to bear upon the arrangement of the electrons in the atoms; but some of it leads to conclusions opposite to those which the remainder suggests.

On the one hand, line-series are discernible in the spectra of elements in the second and the third columns of the Table, and even in those of some others; and from these line-series, systems of Stationary States are deduced which resemble those ascertained for the alkali-metal atoms; and it is natural to extend the same explanation from that case to these, supposing again that each atom consists of a nucleus and a certain number of electrons, all but one of which are tightly bound into a residue, around which the one remaining electron circulates in one or another of various quantized orbits.

On the other hand, the chemical behavior of these elements does not confirm this easy classification of the N electrons of an atom into $(N-1)$ very-tightly-bound electrons and one which is very loosely bound. Thus, the atoms of elements of the second and third columns of the Periodic Table—"alkaline-earth metals" and "earth metals," as they are called—when floating in water as the fragments of molecules of dissolved salts of these elements, are found to be deprived of two and of three electrons, respectively; and the composition of these salts is such as to suggest that the atoms of the other element or elements involved in them have annexed two or three electrons, respectively, from the alkaline-earth atom or from the earth-metal atom. These facts suggest rather that the N electrons of an alkaline-earth atom, or of an earth-metal atom, should be classified into $(N-2)$ or $(N-3)$ very-tightly-bound electrons and two or three which are loosely-bound, respectively. The very tightly bound electrons will be equal in number to, and presumably arranged like, the electrons of the atom of the next preceding inert gas. Henceforth I will reserve the word "kernel" for such a system, and the word "residue" for what is left behind when one electron is separated in fact or in imagination from the atom. Thus these two words will not mean the same thing except in special cases, such as those of the alkali-metal atoms.

Specifically, let us consider the four consecutive elements argon (inert gas, 18th element of the Periodic Table), potassium (alkali metal, 19th element), calcium (alkaline-earth metal, 20th element), and scandium (earth-metal, 21st element).

The evidence from chemistry and from electrolysis impels us to think that the argon atom consists of a nucleus surrounded by (eighteen) electrons tightly bound, in a stable and imperturbable arrangement; that the potassium atom consists of a kernel much like the argon atom, with one additional electron loosely bound and hence

generally far beyond; that the calcium atom consists of the same sort of kernel and two loosely-bound electrons, the scandium atom of the same sort of kernel and three outer electrons.

The Stationary States of the potassium atom have been interpreted as corresponding to various quantized orbits which a single outer electron describes around an unchanging residue; the lines of its spectrum have been attributed to leaps of this electron from one orbit to another, the residue remaining unaltered. There is nothing incompatible between this and the previous conception of the potassium atom.

The Stationary States of the calcium atom resemble, in their arrangement, those of the potassium atom sufficiently to make the same general sort of an explanation desirable,—to make it desirable to suppose that one electron is loosely-bound and remote from the nucleus, the other nineteen tightly-bound and near the nucleus; one loosely-held electron versus nineteen tightly-held ones. But the evidence from chemistry and electrolysis demands two loosely-held electrons versus eighteen tightly-held ones.

One might try to evade the dilemma by supposing that the calcium atom is a sort of three-stage construction, with eighteen electrons congregated in a kernel around the nucleus, a nineteenth far out by comparison with the nucleus, a twentieth far out by comparison with the nineteenth. For interpreting spectra, the residue of the atom would be the kernel or "inert-gas shell" and the nineteenth electron, the valence-electron would be the twentieth. For interpreting chemical data, the residue of the atom would be the inert-gas shell. This conception would rescue the interpretation of the calcium spectrum made after the fashion of the one just expounded for alkali-metal atoms. It would probably demand a larger atom, or a more shrunken kernel, than other data will allow.

Or one might suppose that the nineteenth and the twentieth electron are on the whole about equally remote from the nucleus, and yet it is possible for one of them to change over between any two of a vast system of quantized orbits without greatly affecting the other. There is certain evidence for this conception which I shall presently narrate.

Or one might suppose that the nineteenth and the twentieth electron are a system by themselves, and that each Stationary State corresponds to a particular configuration of this system, so that each line of the spectrum is attributed to a leap not of either electron separately but of both together. This idea seems to be gaining ground rapidly in dealing with atoms composed of a kernel and several outer electrons,

three or four or five or six or seven. The preceding notion might be brought under it as an especial case. If it is accepted the theory of atoms other than the alkali-metal atoms will inevitably be more complex than the theory mentioned for these in section P.

An interesting feature of some of these spectra discloses that the residue of the atom may exist in either of two distinct states. It will be recalled that the energy-values of the Stationary States have been measured from the state of the ionized atom, to which the energy-value zero is assigned. In this fundamental state, one electron and the residue of the atom are completely sundered; and the energy-value of any other Stationary State is the energy required to tear the electron completely out of the atom when the latter is initially in that Stationary State. This definition implies that the state attained when the electron is completely separated from the rest of the atom is determinate and unique. Such must be the case if the atom consists of an invariable nucleus and one electron, as in hydrogen; but if the atom contains several electrons, there is no *a priori* reason for excluding the possibility that there may be several "states of the ionized atom"; in each of these states one electron will be far away, but the residue will have as many different arrangements as there are different states. Extending this idea, one infers that there may be two or more distinct sets of Stationary States for certain elements, each set culminating in a different final configuration of the residue,—that is to say, of the ionized atom.

Several instances of atoms possessing two such distinct families of Stationary States are known; the most noted is probably that of neon, but I will describe the case of calcium, lately interpreted by Russell and Saunders and independently by Wentzel. Two families of terms "primed" and "unprimed," had been identified in the spectrum of this element, and important sequences of each could be followed sufficiently far to make the extrapolation to the limit not too daring. The limits were different, showing that the amount of energy required to separate an electron from an atom initially in its normal state had two values differing from one another by 1.72 equivalent volts. Consequently the residue may remain (it is not necessary to assume that it can long remain) in either of two States differing from one another (when the extra electron is far away) by this amount.

At this point a very significant numerical agreement enters upon the scene. The residue of the calcium atom, the *ionized-calcium* atom, has itself a spectrum which is known, and from which its system of Stationary States has been learned and mapped. Like the systems of Stationary States possessed by neutral atoms, this one includes

s, *p*, *d* and other sequences. The Normal State of the ionized-calcium atom belongs to the *s*-sequence; following the usual custom it may be called the (1, *s*) State. The State of next lowest energy-value, the "next-to-normal" State (so to speak) belongs to the *d*-sequence, and may be called the (3, *d*) State. The energy-difference between the (1, *s*) State and the (3, *d*) State is 1.69 volts. This agrees within the error of the experiments with that value 1.72 equivalent volts, which was found for the energy-difference between the two conditions, in either of which the residue of the calcium atom might be left after the twentieth electron is abstracted. This agreement shows that the extraction of the 20th electron from a calcium atom may leave the residue either in the (1, *s*) State or in the (3, *d*) State.

If now we remember that the ionized-calcium atom is comparable with the potassium atom (and with alkali-metal atoms generally) having as it does eighteen electrons very tightly bound as a kernel around the nucleus and one electron loosely held—then it is reasonable to use the same interpretation of its Stationary States as was expounded in Section P; and to suppose that when the ionized-calcium atom is in the (1, *s*) State that loosely-held electron is revolving in a certain n_1 orbit, and when the atom is in the (3, *d*) State the electron is revolving in a certain n_3 orbit. Thus the extraction of the 20th electron of the calcium atom may be supposed to leave the 19th electron sometimes in the one, sometimes in the other of these two orbits.

We may now inquire whether the 19th electron will always remain in its n_1 orbit, or in its n_3 orbit as the case may be, when the 20th electron reenters the atom, descending from one orbit to another. Here it is necessary to watch one's mental steps very closely; for one is liable to slip into the naive notion of a particular orbit, say for instance a 3_3 orbit, as a fixed and permanent railway-track around which the electron continually runs until something violent derails it. This could not be true unless (to take this special case) the 20th electron had no influence whatever upon the 19th. Were it so, every Stationary State of the one family would differ by the same amount. 1.69 equivalent volts, from the corresponding State of the other family. In fact, the energy-difference between corresponding States varies from one pair to another. This may well be simply because the approach of the 20th electron so alters the forces acting upon the 19th, that its orbit is changed both in geometry and in energy-value, while remaining still identified with the same values of its quantum-numbers. The experiments neither prove nor disprove this; it is commonly accepted as true.

It is a very important fact that the atom may pass from a State of one family to a State of the other,—in terms of the model, that the 19th electron passes from its n_3 orbit to its n_1 orbit, and simultaneously the 20th electron makes some transition or other of its own. The emitted radiation contains the energy resulting from both changes simultaneously, fused together without any discrimination.

R. Building-up of Atoms by "Binding" of Successive Electrons

I next point out that the processes whereby the lines of an optical spectrum are emitted may be regarded, if this theory of the atom is valid, as stages in the gradual formation of an atom. Consider the hydrogen spectrum to begin with; each line is emitted as the atom passes from one Stationary State to another of lower energy-value, the state of least energy being the Normal State of the perfected atom and the state of greatest energy being the condition in which the atom-residue and its electron are torn apart. The various lines of the spectrum correspond to various partial steps along the path from the latter of these states to the former, to various stages of the formation of a hydrogen atom from two separated parts. The specific conception of each Stationary State as a definite orbit of the electron about a nucleus merely reinforces this way of envisaging the process. In the spectra of ionized helium and of neutral helium, we read the testimony of the gradual formation of a helium atom out of a nucleus and two electrons initially quite dissevered. The various lines of the ionized-helium spectrum correspond to different stages in the advance of an electron from the state of freedom to the state of most stable association with a nucleus of charge $2e$, or in Bohr's language, to different stages in the "binding" of an electron by a nucleus of charge $2e$. The various lines of the neutral-helium spectrum correspond to stages in the "binding" of a second electron by a system composed of a nucleus of charge $+ 2e$ and an electron already bound to it. Thus the two spectra of helium testify to two consecutive processes in the upbuilding of a helium atom out of its constituent parts.

The process of building up an atom, by successive adhesions of electrons to an incomplete electron-system surrounding a nucleus—that is to say, the process of building a system of Z electrons around a nucleus bearing the charge Ze, out of a system of $(Z\text{-}b)$ electrons surrounding the nucleus, by consecutively adding b electrons one after the other—evidently occurs very profusely in intense high-current high-voltage discharges in vapours, such as the condensed

spark and above all Millikan's "Vacuum Spark." To take instances from the work of Millikan and Bowen, Paschen, and Fowler: in the spectra of such discharges lines have been identified which belong to atoms for which $Z=14$ and b has the several values 1,2,3,4 (four stages in the building of a silicon atom); and to atoms for which $Z=10+b$ and b has the several values 1,2,3,4,5,6. Many of these spectra of multiply-ionized atoms have not yet been analyzed, but the work is proceeding rapidly. There is reason to hope that within a few years we shall be in possession of interpreted spectra not only of many systems of Z electrons about a nucleus of charge Ze, but also of many systems of fewer than Z electrons about nuclei of charge $+Ze$. This may be highly important, as I will try to show by an illustration. We will consider two consecutive elements of the periodic table; sodium ($Z=11$) and magnesium ($Z=12$).

A Mg atom is imagined as 12 electrons around a nucleus of charge $+12\ e$. It is formed when one electron joins itself to a Mg+ion, which is composed of 11 electrons about a nucleus of charge $+12\ e$. For this process a spectrum is emitted, the so-called arc spectrum of Mg or "MgI" spectrum, which is known and analyzed. It shows that the normal state of the Mg atom is an *s*-state (probably of total quantum-number 3). It is likewise a singlet-and-triplet-spectrum. The first of these facts is taken to mean that the valence-electron, or *twelfth electron* (the reader will see the reason for this usage, the electron in question being the last annexed out of the twelve) of the Mg atom moves in a 3_1 orbit. The second is taken to mean something or other about the residue of the atom, as will be shown in section S.

This residue of the atom is itself formed when one electron joins itself to a Mg++ion, which is a group of 10 electrons about a nucleus of charge $+12\ e$. In this process the so-called spark-spectrum of Mg, or "MgII" spectrum, is emitted. It is known and analyzed. It shows that the normal state of the Mg+ion is an *s*-state (probably of total quantum-number 3). It is a doublet spectrum. The first of these facts is taken to mean that the valence-electron or *eleventh electron* of the Mg+ion, moves in a 3_1 orbit. The second is taken to mean something or other about the residue of the Mg++ion.

A very interesting question now arises: is the Mg+ion actually the same as the residue of the Mg atom? In other words: when a 12th electron is added to the group of 11 electrons about a nucleus of charge $+12\ e$, is the group of eleven left unchanged? If so, we have knowledge about this group from two sources. The character of the MgI spectrum (the fact that it is a singlet-and-triplet spectrum) teaches something about the group, though what it is is far from

clear. The character of the MgII spectrum teaches something about the group, viz., that its eleventh electron moves in a 3_1 orbit. If these two groups are just the same, then the two independently acquired facts about them may be united into a precious correlation. As a matter of fact it is generally assumed that they are nearly if not quite the same. A valuable piece of evidence bearing upon precisely this point, although relating to a different element, was described in the foregoing section.

This suggests that it would be a most desirable achievement to produce the spectra due to groups of $(Z-b)$ electrons congregated about a nucleus of charge $+Ze$, for some value of Z (the higher the better) and all values of b from 0 to $(Z-1)$. Were this done we could almost lay claim to having witnessed the creation of an atom from fundamental particles common to all matter. We could not quite make this claim, since the nucleus of charge $+Ze$ would still remain characteristic of that one kind of atom alone; but we should have made a substantial approach to it. However, there is no immediate prospect of achieving this except for the cases $Z=1$ and $Z=2$ which have already been considered. Our inability to produce the spectrum expected for $Li++$ (i.e. for $Z=3$ and $b=2$) acts as a barrier against utterly tearing down the electron-structures of higher atoms so that they can rebuild themselves before our eyes from the foundations.

The next important question may be introduced in this fashion. Suppose that nothing were known about the spectrum called MgII, therefore nothing about the process of adding an eleventh electron to a group of ten around a nucleus of charge $12\,e$. Knowledge would still be available about the process of adding an eleventh electron to a group of ten about a nucleus of charge $11\,e$; for this is precisely the process which creates the neutral sodium atom out of the Na+ion, and results in the emission of the NaI spectrum or arc spectrum of sodium. This spectrum is a doublet spectrum, and it shows that the normal state of the sodium atom is an s-state, probably of total quantum number 3. This last fact is taken to mean that the eleventh electron in a group of eleven electrons about a nucleus of charge $+11\,e$, is revolving in a 3_1 orbit. Could we have assumed that therefore the eleventh electron, in a group of eleven electrons about a nucleus of charge $+12\,e$, is revolving in a 3_1 orbit? There is no *a priori* certainty of this: but the observations on the MgII spectrum, as we have seen, confirm it (and also that the residue of the Mg+ion is like the residue of the Na atom, in causing the next added electron to produce a spectrum of the doublet type).

Were this generally true we could say that each atom in the periodic table is like the residue of the next atom following it; and that the mth electron in the nth atom is revolving in the same sort of orbit as the outermost electron of the mth atom, for every value of n and for every value of m less than that value of n.

However, it is not always true. To take another specific instance, consider the two elements potassium ($Z=19$) and calcium ($Z=20$). The spectrum KI, which is due to a nineteenth electron joining a group of 18 about a nucleus of charge $+19\,e$, and the spectrum CaII, which is due to a nineteenth electron joining a group of 18 about a nucleus of charge $+20\,e$, are dissimilar. The dissimilarity is not quite so great as to affect the normal states of the two systems, K and Ca+, composed of nuclei of charge $19\,e$ and $20\,e$ each surrounded by 19 electrons; both have as normal state an s-state, apparently of total quantum-number 4; it is inferred that in each, the 19th electron revolves in a n_1 orbit. If we consider, however, the first of the d-states (to which the total quantum-number 3 is commonly assigned), we see that in the KI spectrum it has a much larger energy-value than the Normal State, while in the CaII spectrum it has nearly the same energy-value. A short leap of the imagination leads to the conclusion that if we could examine the spectrum produced by a 19th electron joining a group of 18 about a nucleus of charge $+21\,e$, the d-state in question would have a smaller energy-value than any s-state. In this case it would be the Normal State itself,[3] and we should say that the 19th electron, in a group of 19 surrounding a nucleus of charge Ze, revolves in a n_1 orbit if $Z=19$ or 20, but in a n_3 orbit if $Z=21$.

This system of 19 electrons around a nucleus of charge $21\,e$ is a doubly-ionized scandium atom, Sc++. Its spectrum has not been produced, so that the foregoing sentences are still somewhat speculative. What gives them value is the inference that scandium marks a sort of a breach in the regularity of the Periodic System. For most of the elements in the Periodic System, it can be said that the atom consists of a residue which is like the atom of the preceding element, and an additional electron; and that in its turn this atom resembles the residue of the atom of the element next following. To this the regular periodicity of the properties of the elements is ascribed. But when we reach an element of which the atom has a residue distinctly different from the atom of the foregoing element, then the regular variation of the physical and chemical properties is interrupted. Scandium, as a matter of fact, is the first of a group of

[3] In the First Part of this article the impression may have been left that the Normal State of every atom is an s-state. This is not true; in some known cases the Normal State is a p-state, in others an f-state.

elements, the intrusion of which into the Periodic Table brings about a disruption of the simplicity of its first three rows. There are other such intrusive groups of elements, notably the celebrated groups of the rare earths. It is supposed that wherever such a group commences, there the residue begins to vary from one atom to the next. The spectroscopic evidence is lacking; it is awaited with extreme interest.

The reader will very probably have seen one or more tables of the distribution of electrons in atoms; tables in which it is stated, for instance, that the atom of sodium contains two electrons moving in 1_1 orbits, four in 2_1 orbits, four in 2_2 orbits, and one in a 3_1 orbit; or more succinctly that it contains "two 1_1, four 2_1, four 2_2 and one 3_1 electron." Such tables are built by piecing together bits of evidence, some of which are such as I have described in this section, while others are inferences from X-ray spectra, magnetic properties, or observations of still other kinds. That they are still highly speculative is confirmed by the fact that they are continually being remodeled. If we could produce the spectra corresponding to all the stages of formation of an atom, we should be able to set up a tabulation more reliable than any yet put together. Even then, however, we should be confronted with the question whether the addition of a new electron to a kernel fundamentally alters the distribution of those already there.

Having considered the facts at such length in this section, we are entitled to consider the theory. In the coupled cases of hydrogen and ionized helium it was shown by experiment, and rendered plausible by theory, that the Stationary States of the element with one electron and a double charge on its nucleus correspond exactly to those of the element with one electron and a single charge on its nucleus, and are endowed with fourfold the energy of these latter. This conclusion can be extended to cover the case of a valence-electron circulating in an orbit at a great distance from a kernel composed of $(Z-b)$ electrons and a nucleus bearing the charge $+Ze$. The field due to the kernel will at great distances approximate the field due to a solitary nucleus bearing the charge be. We have seen already that when $b=1$ (so that the total charge on the nucleus balances the total charge of the electrons, valence-electron included) the Stationary States corresponding to orbits for which n and k are large coincide with Stationary States of hydrogen. It follows equally that when $b=2$, the Stationary States for which n and k are large have approximately fourfold the energy of stationary states of hydrogen, and coincide approximately with Stationary States of ionized helium. This is verified by experiment, and so are the corresponding conclusions for the cases $b=3$ and $b=4$.

S. Interpretation of Multiplets

Heretofore in the Third Part of this article I have repeated the procedure adopted in the First Part, simplifying the actual facts by writing as though the Stationary States of each atom were arranged in sequences of individual terms, each sequence being distinguished by a particular value of the Azimuthal Quantum Number. Here as there, it finally becomes necessary to concede the complexity of the facts, and recognize that the sequences in question are sequences not of individual terms, but of groups of terms. Thus for instance the sodium atom possesses a *p*-sequence, not of individual terms but of pairs of terms—a pair $2p_1$ and $2p_2$, then a pair $3p_1$ and $3p_2$, then a multitude of other pairs. For another instance, the mercury atom exhibits a *p*-sequence not of individual terms but of triads of terms—a triad $2p_1$ and $2p_2$ and $2p_3$, then a triad $3p_1$ and $3p_2$ and $3p_3$, and then a procession of other triads. These sequences are collected into systems: an *s*-sequence and a *p*-sequence and a *d*-sequence and several more constitute a system. There are singlet systems and doublet systems and triplet systems and systems of still higher *multiplicity*; and each kind of system is distinguished by a certain manner of grouping of the terms which form its various sequences. Noteworthy and peculiar laws govern these groupings; in a doublet system, for instance, the *s*-sequence consists of individual terms, but all the others consist of pairs of terms; in a quartet system, the *s*-sequence is made up of single terms, the *p*-sequence of triads of terms, the remaining sequences of groups of four terms each. From the First Part of this article I reprint a Table showing how the terms are grouped in systems of all multiplicities from the singlet to the octet. The numbers opposite the name of each system and under the letters of the various sequences show how many terms belong to each group in the various sequences of that system.

TABLE I

Name of System	*s*	*p*	*d*	*f*	*f'*	*f''*
Singlet	1	1	1	1	1	1
Doublet	1	2	2	2	2	2
Triplet	1	3	3	3	3	3
Quartet	1	3	4	4	4	4
Quintet	1	3	5	5	5	5
Sextet	1	3	5	6	6	6
Septet	1	3	5	7	7	7
Octet	1	3	5	7	8	8

Each atom possesses one or more such systems of Stationary States; and the particular types which an element displays depend in a definite and fairly clear manner upon the position of the element in the Periodic Table, being in fact one of the most distinctive of the periodically-varying qualities. Each atom with an even number of electrons exhibits systems which are all of odd multiplicity, and each atom with an odd number of electrons exhibits systems which are all of even multiplicity; thus magnesium, with twelve electrons, has a singlet system and a triplet system, while sodium and once-ionized magnesium, each with eleven electrons, have each a doublet system, and neon with ten has a singlet, a quintet and two triplet systems.[4] Remembering what was said about the consecutive binding of electrons, it will be noticed that these facts show a regular difference between the binding of the Nth electron when N is odd and the binding of the Nth electron when N is even. Otherwise expressed, they show that a kernel of N electrons treats an oncoming member in one or another of two distinctive ways, according as N is even or odd. The influence of magnetic fields on spectra likewise shows that this complexity of the Stationary States is a quality not negligible, but primary.

The features of the atom-model hitherto described must be supplemented with some new one if it is to cope with such facts as these. We have represented (for example) the sodium atom in its $2p$ state by a "valence-electron" cruising with angular momentum $2(h/2\pi)$ in an orbit around a "kernel" composed of ten electrons and a nucleus. But there are two such states instead of one; if the angular momentum of the valence-electron is to be equal to $2(h/2\pi)$ for each of these, some other not yet mentioned feature of the atom must discriminate the two. One might, of course, again proceed as we did in discussing the "primed terms," by assuming that the kernel of the atom is in one condition when the atom is in the $2p_1$ state, and in another slightly different condition when the atom is in the $2p_2$ state. This would probably entail as many different conditions of the kernel as there are pairs of terms in the sodium spectrum—a great number, and yet small in comparison with the multitude which would be required by other atoms; yet such may be the eventual theory. However, it is possible to construct for these facts an atom-model out of two revolving parts, whereby different Stationary States of a group are represented not by varying the condition of either part separately,

[4] Hydrogen and ionized helium are not included under this rule. Helium shows a singlet and a doublet system together, a combination which violates the rule as stated, unless the doublet system is really a triplet system in which two states of each triad are too close together to be distinguished.

but by varying the relative orientation of the two. Although this theory has not been harmonized with those which I have hitherto recited, it is competent in its own field; and for that reason I present it.

We will imagine that the atom is represented by a combination of two flywheels, two whirling objects, endowed each with angular momentum. These angular momenta are vectors, pointing along the directions of the axes of rotation of the respective flywheels, and having certain magnitudes. I will designate them temporarily as P_V and P_R, each symbol standing for a vector generally and also (when in an equation) for its magnitude. The angular momentum of the entire atom, which is necessarily constant in magnitude and in direction so long as the atom is left to itself, is the resultant of P_V and P_R; a vector, pointing along the direction of the so-called "invariable axis" of the atom. I designate it by P_A. The following equation shows the relation between the magnitudes of these three angular momenta and the angle Θ between the two first-named, the angle which describes the relative orientation of the axes of rotation of the two flywheels:

$$P_A^2 = P_V^2 + P_R^2 + 2P_V P_R \cos \Theta \qquad (6)$$

Remembering the successes which in dealing with the spectrum of hydrogen have resulted from assuming that the angular momentum of the entire atom is constrained to take only such values as are integer multiples $Jh/2\pi$ of the quantity $h/2\pi$, we make the same assumption here. We further make the same assumption for each of the flywheels separately; the magnitudes of the angular momenta P_V and P_R are supposed to take only such values $Vh/2\pi$ and $Rh/2\pi$ as are integer multiples of the same quantity $h/2\pi$.[5] These particular assumptions, frankly, are foredoomed to failure; but the failure will be instructive.

Making all these assumptions together, we see that in effect we have laid constraints upon the angle Θ which measures the relative orientation of the two flywheels. For if P_V is an integer multiple of $h/2\pi$, and P_R is an integer multiple of $h/2\pi$, then P_A which is fully determined by equation (6) cannot be an integer multiple of $h/2\pi$ unless Θ is very specially adjusted. To illustrate this by an instance (which will be clearer if the reader will work it out with arrows on a sheet of paper): if P_V and P_R are each equal to the fundamental quantity $h/2\pi$, and if P_A must itself be an integer multiple of $h/2\pi$: then cos Θ must take only the values, $+1$, $-\frac{1}{2}$, -1, which yield the

[5] All that is actually being assumed is, that P_V and P_R and P_A are all integer multiples of a common unit; nothing in this section will indicate either $h/2\pi$ or any other value as the precise amount of that common unit.

values O, $2\pi/3$, π for Θ, which yield the values $2h/2\pi$, $h/2\pi$, 0 for P_A. Any other integer values for $P_A/(h/2\pi)$ are unattainable by any value of Θ whatsoever; any value of Θ not among these three would yield a value for P_A not an integer multiple of $h/2\pi$, which is contrary to the assumptions. Thus, the assumptions that the atom is a conjunction of two whirling parts, and that the atom altogether and each of its two parts separately whirl with angular momenta which are constrained to be integer multiples of a common factor—these assumptions lead to the conclusion that the relative inclination of the two revolving parts is constrained to take one or another of a strictly limited set of values.

This essentially is the model devised by Landé to account for the complexity of the Stationary States. The several Stationary States which form a group belonging to a sequence—in other words, which share a common value of n and a common value of k, like the $2p_1$ and $2p_2$ states of sodium or the $3d_1$, $3d_2$, $3d_3$ states of mercury—are supposed to resemble one another in this, that each of the whirling parts separately has the same angular momentum in every case; and to differ from one another in this, that in the several cases the two whirling parts are differently inclined to one another, so that the angular momentum of the entire atom differs from one state to the next. The different Stationary States which share common values of n and k are supposed to correspond to different orientations of the two parts of the atom and to different values of its angular momentum.

I will now no longer disguise the fact that these whirling parts are, or at any rate have been, supposed to be precisely the valence-electron and the residue. To the former we should therefore assign these values for the angular momentum P_V: the value $h/2\pi$ for every state belonging to an s-sequence, the value $2h/2\pi$ for every p-state, $3h/2\pi$ for every d-state, and so on. Then to the angular momentum P_R of the residue we should assign a suitable constant value; a "suitable" value in this case being such a one, as would yield the proper grouping of terms in the various sequences of the system which the atom under consideration is known to have. Thus, for an atom-model to represent sodium with its doublet system we require a value for the angular momentum of the residue, such as will yield one permitted orientation when the atom is in an s-state ($P_V = h/2\pi$), and two when it is in any state for which $P_V = kh/2\pi$ and k is any integer greater than unity.

No such value can be found. The value $P_R = h/2\pi$ will not do; for, as was shown in the illustrative instance a couple of pages back, it yields three permitted orientations when $P_V = h/2\pi$, and (as can easily be shown) three for each and every other value of P_V which

is an integer multiple of $h/2\pi$. Thus it would form an adequate model for a system of Stationary States in which every group of terms in every sequence was a triad; but this is not a doublet system, nor even a triplet system, nor any other observed system whatever. To make this long story short; it is impossible to simulate any of the eight groupings of terms set forth in the eight lines of Table I by assuming that P_V, P_R and P_A are all integer multiples of $h/2\pi$ (or of any other common factor).

It is in fact necessary to put P_V equal, not to $h/2\pi$ and to $2h/2\pi$ and to $3h/2\pi$, but to $\frac{1}{2}(h/2\pi)$ and to $\frac{3}{2}(h/2\pi)$ and to $\frac{5}{2}(h/2\pi)$, for the *s* and *p* and *d* states, respectively. This use of "half quantum numbers" makes it possible to produce an adequate model for an atom possessed of a doublet system, by assuming that the angular momentum P_R of its residue is always $h/2\pi$, and that its two whirling parts must always be so inclined to one another that the angular momentum of the entire atom is an integer multiple of $h/2\pi$.

For (to work out one example, and one only) when we make $P_R=h/2\pi$ and $P_V=\frac{1}{2}(h/2\pi)$, then the greatest possible resultant that can be obtained by combining these vectorially is $\frac{3}{2}(h/2\pi)$ and the least possible one is $\frac{1}{2}(h/2\pi)$; these two extreme values being attained when the two component vectors are parallel and when they are anti-parallel,[6] respectively. If we permit for the resultant only such values as are integer multiples of $h/2\pi$, then there is only *one* that is permitted: the value $h/2\pi$—for this is the only such value lying within the possible range. Next, put $P_R=h/2\pi$ and $P_V=\frac{3}{2}(h/2\pi)$. All possible values of the resultant lie between $\frac{5}{2}(h/2\pi)$ and $\frac{1}{2}(h/2\pi)$; within this range there are *two* of the integer multiples of $h/2\pi$ which are the sole permitted ones. Next, put $P_R=h/2\pi$ and $P_V=\frac{5}{2}(h/2\pi)$. All possible values of the resultant lie between $\frac{7}{2}(h/2\pi)$ and $\frac{3}{2}(h/2\pi)$, and this range again includes *two* permitted values. Thus the model describes properly the grouping of the terms in a doublet system. I leave it to the reader to show that by putting $P_R=2h/2\pi$, or $3h/2\pi$, or $4h/2\pi$, and treating P_V as in this foregoing case, he can reproduce the groupings of terms in quartet, or sextet, or octet systems, respectively, as Table I, describes them.[7]

A more drastic use of "half quantum numbers" is required to obtain an adequate model for atoms showing singlet and triplet and other

[6] This convenient word is used in German to describe vectors pointing in opposite senses along the same direction.

[7] The diagrams with arrows, offered by Sommerfeld in the fourth edition of his classic book, are very helpful in studying these models. Incidentally Sommerfeld's alternative way of arriving at the groupings of multiplet terms by compounding vectors is instructive.

systems of odd multiplicity. Thus to produce a singlet system it is necessary to put $P_R = \frac{1}{2}(h/2\pi)$ always; to set P_V equal to $\frac{1}{2}(h/2\pi)$ for all S-states, to $\frac{3}{2}(h/2\pi)$ for all P-states, and so forth; and to suppose that the two whirling parts of the atom are constrained to take only such relative orientations as yield values for P_A, the angular momentum of the entire atom, which are odd integer multiples of $\frac{1}{2}(h/2\pi)$. It is easy to see that there is but one such orientation for an s-state, one for a p-state, and one for any other kind of state. To produce a triplet, or a quintet, or a septet system, it is necessary to put $P_R = \frac{3}{2}(h/2\pi)$, or $\frac{5}{2}(h/2\pi)$, or $\frac{7}{2}(h/2\pi)$, respectively; and to retain the just-stated assumptions about P_V and P_A.

The question whether these models have any intrinsic truthfulness has now become acute. If there is any doctrine in contemporary atomic theory which appears to be multiply tested and approved, it is surely the doctrine that the angular momentum of the valence-electron is always an integer multiple of $h/2\pi$. Yet in this passage I have spoken as if this principle had been indifferently and casually discarded, and replaced by a new principle to the effect that the angular momentum of the valence-electron is always an odd-integer multiple of $\frac{1}{2}\ h/2\pi$. It is hard to evade or mitigate this arrant contradiction.

A way out may possibly be found by suggesting that the partition of an atom into "residue" and "valence-electron," while appropriate when calculating energy-values by the method mentioned in Section P, is not appropriate in this instance; that the two whirling parts of the atom are respectively a system composed of a part of the residue, and a system composed of the rest of the residue and the valence-electron. This seems most admissible for such an atom as magnesium, consisting as it is supposed of what I have called a "kernel," and two additional electrons outside. The two whirling parts may be the kernel rotating as a unit, and the pair of outer electrons also rotating as a unit. It may be profitable to push the analysis even further, and to consider the two outer electrons each as an entity possessed of angular momentum, their two angular momenta combining with one another in such a fashion as I have lately described for the two parts of the atom; this resultant angular momentum of the two may then figure as the P_V employed in constructing the atom-model. There are decided possibilities in this way of thinking; but it is doubtful whether the difficulty about half-quantum-numbers can ever quite be removed.[8]

[8] An unfortunate feature of Landé's model in its original form is that it requires us to believe that the residue of each atom is different from the completed preceding atom. For instance since Mg has a singlet and a triplet system, its residue must have sometimes $P_R = \frac{1}{2}(h/2\pi)$ and sometimes $P_R = \frac{3}{2}\ h/2\pi$; whereas for the Na atom in its normal state $P_A = h/2\pi$ by the theory.

It may be recalled from the First Part of this article that the different Stationary States of a group, sharing a common value of n and a common value of k, are distinguished from one another by having different values of a numeral which was designated by j and called the Inner Quantum-number. It was so chosen that the only transitions which occur are those in which j change by one unit, or not at all; while transitions between two states, in each of which $j=0$, are likewise missing. This numeral is correlated with the angular momentum P_A of the entire atom, in the theory here outlined. For systems of even multiplicity, P_A is equal to $jh/2\pi$; for systems of odd multiplicity, P_A is equal to $(j+\frac{1}{2})\,h/2\pi$.

The various Stationary States of a group differ slightly in energy—otherwise, of course, they would never have been discerned. The energy-value of an atom must be conceived therefore as depending not merely upon n and k, not merely on the rates at which the two whirling parts are separately spinning, but likewise upon their mutual orientation and hence upon j. In this theory, the dependence of energy upon orientation must be postulated outright. We shall presently meet with a case in which the dependence of energy upon orientation can be foreseen, even in detail.

It appears from all these speculations, that a transition between two Stationary States is no longer to be conceived merely as a simple leap of an electron from one geometrically-definite orbit into another. A leap is indeed supposed to occur, but it is accompanied by a turning-inward or a turning-outward of the axes of rotation of the two spinning parts of the atom. The radiation which comes forth is a joint product of these two processes, in which however no features of either separately appear; only the net change in the energy of the atom, the algebraic sum of the energy-changes due to each process separately, is radiated as a single fused unit. Nature does not make the separation which our imaginations make.

T. Magnetic Properties of Atoms

Having used an orientation-theory to interpret the complexity of the Stationary States, we will now consider an orientation-theory developed to account for the effect of a magnetic field upon the Stationary States. There, it was supposed that the various States belonging to a single group are distinguished by various orientations of two spinning portions of an atom, relatively to one another. Here, it will be supposed that the various States which replace each individual State, when a magnetic field acts upon the atom, are distinguished

by various orientations of the spinning atom relatively to the field. It will presently be seen that the evidence for the orientation-theory is much more abundant and more nearly direct, in this case of magnetically-excited Stationary States, than in that former case of multiplets. This case in fact was the earliest to which an orientation-theory was applied; but for it, some quite different form of theory might have been developed for multiplets. Even here the data and the theory are not entirely concordant; but the concordance is so extensive, that the discord is sharply localized and identifiable.

From the Second Part of this article (Section L) I quote the principle that an electron (of mass μ) revolving in an orbit with angular momentum P is equivalent to a magnet of which the magnetic moment M is proportional to P, being

$$M = eP/2\mu c \tag{7}$$

Both P and M are vectors normal to the plane of the orbit and hence parallel to each other. If several electrons are revolving in divers plane orbits about the same nucleus, their separate angular momenta may be summed vectorially into a vector which is the angular momentum of the entire system, and their separate magnetic moments may likewise be summed vectorially into a vector which is the magnetic moment of the entire system; and these two summation-vectors will be parallel to one another, and related by the foregoing equation. Hence a rigidly-connected revolving framework of electrons—if such a thing there be—may be treated like a single electron, insofar as the ratio of magnetic moment to angular momentum is concerned. Wherever in the course of this article we have envisaged electrons, kernels, or atoms revolving with angular momenta prescribed as integer multiples of $h/2\pi$ or of $\frac{1}{2}\, h/2\pi$, we might have imagined these as magnets with magnetic moments prescribed as integer multiples of $eh/4\pi\mu c$ or of $\frac{1}{2}\, eh/4\pi\mu c$.[9] This is not necessary; though the relation between angular momentum and magnetic moment is derived directly from an equation valid for perceptible electric currents, it might not be true for individual electrons. Nevertheless we shall arrive at striking results, by supposing that it is.

When a magnetic field is applied to a multitude of radiating atoms, most of the lines of their spectrum are replaced by groups of several lines each, or "split up" into several components, as the phrase is. This signifies that each of the Stationary States of each atom is apparently replaced by several. One may infer that when an atom is

[9] The quantity $eh/4\pi\mu c$, the presumptive magnetic moment of an electron circulating in an orbit of angular momentum $h/2\pi$, is known as the *Bohr magneton*.

introduced into a magnetic field, each of its Stationary States is modified into one or another of several new States, differentiated from one another and from the original State to a small but appreciable extent. This might arise from some distortion or internal alteration of the atom by the field; and it will probably be necessary to adopt this view in some cases. But there is also a simpler effect which the magnetic field may have upon the apparent energy-values of the Stationary States, an effect not involving any deformation of the atom by the field—to wit, an orientation-effect similar to that which was assumed to account for multiplets. This we proceed to examine.

If an atom which is a magnet is floating in a magnetic field, it experiences a torque which tends to orient it parallel with the field. By saying that an atom is parallel or oblique to the field, I mean that the magnetic moment of the atom and therefore also its angular momentum, are directed parallel or obliquely to the field; and this usage will be maintained. Owing to this torque it is endowed with energy due to the field, in addition to its own intrinsic energy; this additional energy, which depends upon the inclination of the atom to the field, I shall call its *extra magnetic energy*. If the atom turns in the field, the amount of its extra magnetic energy changes; and if its magnetic moment suddenly changes, its extra magnetic energy also changes unless it simultaneously turns by just the right amount to compensate the change. If at the moment of passing over from one of its Stationary States to another, its inclination or its magnetic moment or both are changed; the amount of magnetic energy which it gains or loses will be added (or subtracted, as the case may be) to the amount of energy which it gains or loses because of the transition. The frequency of the radiation sent out or taken in by the atom will be equal to $1/h$ times the sum of two energy-changes of distinct kinds—not, as in the absence of magnetic field, to $1/h$ times the energy-difference between the two Stationary States alone. Thus the effect of a magnetic field upon spectrum lines might be ascribed, not to any deformation of the atom by the field, but to changes in the orientations or in the magnetic moments of the atoms occurring at the instants when they make their transitions. The question for us now is, whether the actual details of the observed effect can be interpreted in this manner.

Expressing the foregoing statements in formulae, in which M denotes the magnetic moment of an atom, H the magnetic field, and α the inclination of the atom to the field, we have for the torque which the field exerts upon the atom

$$T = MH \sin \alpha \qquad (8)$$

and for the "extra magnetic energy" of the atom due to the field

$$\Delta U = -MH \cos \alpha \tag{9}$$

In this last expression it is tacitly assumed that the extra magnetic energy is zero when the atom is oriented crosswise (at right angles) to the field. This is not an arbitrary, but a quite essential convention, justified from the atom-model.[10] Suppose now that the atom passes between two stationary states S' and S'', in which its internal energy, its magnetic moment and its inclination are denoted by U', M', α' and U'', M'' and α'', respectively. Were there no magnetic field, the frequency radiated would be

$$\nu_o = (U'' - U')/h \tag{10}$$

but owing to the field, the frequency radiated is

$$\nu_H = \nu_o + \Delta\nu = (U'' - U')/h + H\,(M' \cos \alpha' - M'' \cos \alpha'')/h \tag{11}$$

the term $\Delta\nu$ representing the displacement of the line by the field. The question is, whether this term can be equated to the observed displacements.

Consider the most tractable cases, those in which the so-called "normal Zeeman effect" is observed. In these cases a line of frequency ν_o is replaced by three, of which the frequencies are

$$\nu_o + \omega H,\ \nu_o,\ \nu_c - \omega H \tag{12}$$

corresponding to three values for the displacement $\Delta\nu$, which are expressed by

$$\Delta\nu = +\omega H,\ O,\ -\omega H \tag{13}$$

The quantity ω occurring in these expressions is a specific numerical constant. Comparing these with the expressions for $\Delta\nu$ in (11), we see that if our model is to be used to interpret the observations, then for the first of the three observed lines $M' \cos \alpha'$ must be greater than $M'' \cos \alpha''$ by the amount $h\,\omega$; for the second, $M' \cos \alpha'$ must be equal to $M'' \cos \alpha''$; for the third, $M' \cos \alpha'$ must be less than $M'' \cos \alpha''$ by the amount ωh.

Another way of putting these statements is, that in order to interpret the normal Zeeman effect in this manner it must be supposed that whenever a transition occurs, the projection of the magnetic moment

[10] The action of the magnetic field upon the revolving electron imparts to it an extra angular velocity about the direction of the field (the Larmor precession) and hence an extra kinetic energy which (to first order of approximation) is proportional to $-\cos \alpha$ and is zero when $\alpha = \pi/2$. This extra kinetic energy is the extra magnetic energy ΔU. It is profitable to derive the entire theory in this manner.

upon the direction of the field—for this is precisely what $M \cos \alpha$ is—either does not change at all or else changes by $\pm\omega h$. Sometimes it acts in the first of these ways, sometimes in the second, sometimes in the third; but never in any other.

This would result, if the behavior of the atom floating in the magnetic field were governed by two rules; *first*, that it may orient itself only in certain "permitted" directions such that $M \cos \alpha$, the projection of its magnetic moment upon the field-direction, assumes "permitted" values which are integer multiples of ωh; second, that whenever a transition occurs $M \cos \alpha$ either retains the value which it had initially, or else passes to one or the other of the two adjacent permitted values.

The first of these rules is stated more rigorously than is quite necessary; all that is required is to say that $M \cos \alpha$ is permitted to take only such values as belong to an equally-spaced series with intervals equal to ωh. The second rule is necessary.

The theory of the normal Zeeman effect is simply, that the atom does behave according to these rules. Radiation of the frequency ν_o occurs, either when the magnetic moment of the atom does not change and the atom does not turn, or when the magnetic moment changes and simultaneously the atom turns just so as to keep the projection of the magnetic moment on the field-direction constant. We shall later see that the latter of these two alternatives is the accepted one. It must be supposed that the atom, so to speak, capsizes when it emits the frequency ν_o while floating in a magnetic field; it flops over at the same moment as it passes from one stationary state to another. Radiation of the frequency $\nu_o+\Delta\nu$ or of the frequency $\nu_o-\Delta\nu$ occurs, as we shall see, when the magnetic moment of the atom changes; in some cases the atom capsizes during the process, in others it does not.

I now translate the foregoing rules from the language of magnetic moments to the language of angular momenta. The first rule is, that the atom may orient itself only in certain permitted directions such that $P \cos \alpha$, the projection of the angular momentum upon the direction of the magnetic field, assumes permitted values which are consecutively spaced at intervals of $(2\mu c/e)\omega h$.

Now it is a fact of experience, that in the cases of the normal Zeeman effect,

$$\omega = e/4\pi\mu c. \tag{14}$$

The rule therefore reads, that *the projection of the angular momentum of the atom upon the direction of the magnetic field is constrained to take certain permitted values, spaced at intervals of* $h/2\pi$.

We have supposed, in dealing with multiplets, sometimes that the angular momentum of the entire atom is constrained to take such values as are integer values of $h/2\pi$, and sometimes that it is constrained to take such values as are odd-integer multiples of $\frac{1}{2}$ $(h/2\pi)$.[11] In either case the permitted values of the angular momentum are spaced at equal intervals; and as the rule for the component of the angular momentum along the direction of the field bears the form which it does, we may well suppose that something in the order of nature constrains both the angular momentum and its projection to accept only values which form a sequence spaced always at that curious interval $h/2\pi$.

The total number of permitted orientations will obviously be limited by the actual magnitude P of the angular momentum. This being supposed always to be an integer multiple of $\frac{1}{2}$ $h/2\pi$, let it be written $P=2J(\frac{1}{2}\ h/2\pi)$. The permitted orientations are those which yield a series of values for the projection $P\cos\alpha$ spaced at intervals $h/2\pi$; let these be written

$$P\cos a=A_o,\ A_o-h/2\pi,\ A_o-2h/2\pi,\ \ldots\ .\ A_o-mh/2\pi \qquad (15)$$

Nothing in the experiments thus far described gives the least notion of the value which should be assigned to A_o. All we know at present is that A_o cannot exceed P and that $(A_o-mh/2\pi)$ cannot be algebraically less than $-P$. Suppose in the first place that $A_o=P$; that is to say, that the atom may orient itself with its axis parallel to the magnetic field. Then the permitted orientations are those which yield this series of values of the projection:

$$\begin{aligned} P\cos\alpha=P,\ P-h/2\pi,\ P-2h/2\pi,\ \ldots\ldots\ .\ P-mh/2\pi \qquad (16)\\ =2J(\tfrac{1}{2}\ h/2\pi),\ (2J-1)(\tfrac{1}{2}\ h/2\pi),\ (2J-2)(\tfrac{1}{2}\ h/2\pi,\ \ldots\ .\ ,\ 0 \end{aligned}$$

of which there are $(2J+1)$ in all. On the other hand, it may be that the atom is prevented from orienting itself parallel to the field; that the least permitted angle between the axis of the atom and the direction of the field is some angle yielding a projection A_o intermediate between P and $(P-h/2\pi)$. Then there are $2J$ permitted orientations altogether.

Summarizing the results of this last paragraph: if the angular momentum of the atom is an integer multiple $2J(\frac{1}{2}\ h/2\pi)$ of the fundamental unit $\frac{1}{2}$ $(h/2\pi)$, then according to the orientation theory

[11] It was remarked at the beginning of Section S that the evidence to be presented in that Section would support neither $h/2\pi$ nor any other particular numerical value for the fundamental unit of angular momentum; here, however, we have evidence for that value.

the atom is permitted to take either $(2J+1)$ or $2J$ distinct orientations in the field; the former number if it is, the latter if it is not permitted to set itself quite parallel to the field.

It will now be shown that these are by no means idle speculations; they bear directly upon certain facts accessible to observation. Before bringing up these facts it is necessary to abandon the policy of speaking exclusively about the "normal" Zeeman effect. This "normal" effect received its adjective because it agrees so excellently with the original theory devised years before quanta were dreamt of to explain the effect of magnetic field upon spectra. It is essentially because of this agreement that it is possible to develop the contemporary theory of the "normal" effect in a perfectly deductive fashion, using no new assumptions beyond those general ones of the orientation-theory. Most spectrum lines, however, are affected by a magnetic field in ways not compatible with the original theory; which is a consequence of the fact that the set of new Stationary States, whereby a magnetic field supplants each original Stationary State, in most cases does not conform to the laws previously set forth.

The laws to which it generally does conform were read from the spectra by Landé. The one feature in which the foregoing theory quite agrees with these laws is its prediction of the total number of Stationary States. A Stationary State for which the angular momentum of the atom is determined, by virtue of the theory of multiplets which filled the preceding section of this article, as being $2J\,(\frac{1}{2}h/2\pi)$, is actually found to be supplanted, when a magnetic field is impressed upon the atom, by $2J$ new Stationary States. This is in agreement with one of the two alternative predictions made a few paragraphs *supra*; to wit, with the prediction derived from the assumption that the atom cannot set itself quite parallel to the field. This agreement between the orientation-theory of multiplets and the orientation-theory of Zeeman effect considerably strengthens both.

The several Stationary States replacing a given original State are always equally spaced; but the spacing differs in amount from the value ωIIh or $eIIh/4\pi\mu c$ exhibited when the normal Zeeman effect occurs, and which we found it possible to deduce from the simple orientation-theory. The difference is this, that the actual spacing is a multiple of the value ωIIh by a factor g (generally lying between $\frac{1}{2}$ and 2) which depends upon the original State:

$$\Delta U = g\omega IIh \qquad (17)$$

The only ways hitherto used to accommodate the atom-model to this surprising and inconvenient factor g are tantamount to assuming that it enters into the relation between angular momentum M and magnetic moment P which was derived in Section L and written down here as equation (7); which relation is accordingly modified without discoverable reason into

$$M/P = g\, e/2\, \mu c \tag{18}$$

a very unsatisfying procedure. Lande found it possible to mitigate this process somewhat and at the same time produce a partial explanation of the formula quoted in the First Part of this article, whereby g is related to the factors K, R and J which, in the atom-model of the two whirling parts, measure the angular momenta of valence-electron and residue and entire atom respectively in terms of the common unit $h/2\pi$. This explanation involved the postulate that $g=1$ for the valence-electron and $g=2$ for the residue. It would therefore be necessary to justify, or to postulate without justification, not a multitude of such relations as (16) with a multitude of unforeseen values of g, but only a single such relation with a single unforeseen value of g. This is bad enough, but not so bad as if it were inevitable to assume that M/P may have a dozen different values in different cases.

It is now the occasion to recur to the extraordinary experiments of Gerlach which disclose the magnetic moments of individual atoms and verify the supposition that certain orientations are permitted and others inhibited. These experiments having already twice been mentioned in this series of articles, I shall spend no more space upon the method than is necessary to say that atoms in a narrow stream are sent flying across an intense magnetic field with a strong field gradient, by which they are drawn aside. Were the atoms tiny magnets oriented randomwise in all directions, the beam would be broadened into a fan; one edge of the fan would be the path of atoms oriented parallel to the field, the other edge would be the trajectory of atoms oriented anti-parallel to the field, while the space between the edges would be filled by the orbits of atoms pointed obliquely to the field. Actually Gerlach observed not the whole fan, but two or several separate diverging pencils of atoms, and between them vacant regions traversed by none. Certain orientations of atoms to field were unrepresented in the beam. Here for the first time there is direct evidence of discrete Stationary States, of quantum permissions and quantum inhibitions, not deduced from observations upon transi-

tions but drawn forthright from viewing atoms in equilibrium in their Normal States.

When from the diverging pencils one proceeds to determine the orientations of the atoms and their magnetic moments, one is confused by a possibility made clear in the foregoing pages, but unsuspected at the time when the first of these experiments were performed. I illustrate with the case of silver, the atoms of which flock into two diverging pencils with a quite vacant space between. At first it was naturally supposed that one pencil consists of atoms oriented parallel, the other of atoms oriented anti-parallel to the field. The deflections of the two pencils are such, that if this assumption is true then the numerical value of the magnetic moment of the silver atom agrees within the error of experiment with the value of $eh/4\pi\mu c$—agrees, therefore, with the notions that the angular momentum of the silver atom in its normal state is $h/2\pi$ and that the magnetic moment stands in the right and proper ratio $e/2\mu c$ to the angular momentum. The data were supposed to prove these notions. They also agree, however, with the suppositions that one pencil consists of atoms inclined at 60° to the field and the other of atoms inclined at 120°; in which case the magnetic moment of the silver atom would be $2eh/4\pi\mu c$, suggesting that the ratio of magnetic moment to angular momentum has twice the right and proper value. This inextricable tangling of the effect of orientation with the effect of magnetic moment makes it impracticable to deduce quite so much from the data as was at first thought possible; but plenty still remains. It is found that copper and gold behave like silver, as was to be expected from their positions in the Periodic Table. It is found that lead atoms, and (most surprising of all!) *iron* atoms are not deflected at all; so that either the magnetic moments of their parts balance one another completely, or else they all orient themselves crosswise to the field. Nickel, on the other hand, behaves as though its atoms had each a magnetic moment surpassing $2eh/4\pi\mu c$, while thallium responds as though that of its atoms were much less than $eh/4\pi\mu c$. Finally—lest the results seem too gratifying—it is found that bismuth atoms are deflected in a manner quite unforeseeable.

There is not time nor space to speak of the other method for determining the magnetic moments of atoms, by measuring the susceptibilities of great quantities of them in gases or solutions; but the measurements so made are also very helpful in determining the magnetic moments of various atoms and ions—various groupings, that is to say, of electrons around nuclei.[12] All such data are of im-

[12] For the status of such measurements in 1923, the first of this series of articles may be consulted. (This Journal, September, 1923.)

mense value, and no theory of the atom can be spared from the demand that it confront them and account for them.

U. Interpretation of X-Ray Spectra

By the term "X-ray" the reader may understand any radiation of which the frequency ν is so high, that the energy $h\nu$ of a single quantum is several times as great as the energy required to remove the most-easily-detached electron from an atom; greater, for instance, than 100 equivalent volts, so that the wavelength of the radiation is less than some 125 Angstrom units. The emission or the absorption of such radiation by an atom involves too great an energy-change to be attributed merely to a displacement of the valence-electron or even to combined displacements of the valence-electron and one or two others. This definition leaves a sort of "twilight zone" of radiations having frequencies somewhat but not much greater than $1/h$ times the ionizing-potential of an atom. Little is known about such radiations, and in this place they will not be considered.

The absorption of an X-ray quantum by an atom results in the extrusion of an electron from the atom. The emission of an X-ray quantum results from the passage of an electron within the residue of the atom from some original situation to the situation vacated by the extruded electron, or else into a situation vacated by an electron which itself has moved elsewhere within the atom. These statements contain the theory of the vast amount of data piled up by observations upon the emission and absorption of X-rays by matter.

To express the same statements rather differently: X-ray absorption-spectra and X-ray emission-spectra reveal, when analyzed for Stationary States in the manner used in analyzing optical spectra, that each atom with several or many electrons has a considerable number of Stationary States, distinguished from those we have heretofore discovered in that *each of them involves the absence of one electron from the atom.* Each of them is therefore strictly an "ionized-atom state," and yet there are several of them with extremely different energy-values. This signifies that the extraction of an electron from an atom rich in electrons may leave the residue in any one of several distinct conditions. These distinct conditions are the distinct Stationary States, transitions between which are responsible for X-ray spectra. Owing to this striking difference between the Stationary States hitherto described and these latter, I shall refer to these as the "X-ray Stationary States"—not that this name is a particularly good one.

Absorption of an X-ray quantum by an atom, then, results in a transition of the atom from its normal state to one of the "X-ray Stationary States." Emission of an X-ray quantum by an atom results from a transition of the atom from one into another of its "X-ray Stationary States"—a transition which begins in a condition in which the atom lacks one electron, and ends in another condition in which the atom lacks one electron. To take instances: radiation of an adequate frequency falling upon an atom in its normal state may put it over into an X-ray Stationary State known as the L_{II} state, an electron being extruded. Radiation of an adequate frequency (a higher frequency will be required) falling upon a similar atom in its normal state may put it over into another X-ray Stationary State of higher energy, known as the K state, an electron being extruded. The atom in the K state may then spontaneously pass over into the L_{II} state, emitting a radiation belonging to the X-ray emission-spectrum, its frequency being $1/h$ times the energy-difference between the K state and the L_{II} state. Later the atom may pass into still another state, such as the M_I state, by emitting radiation of some frequency ν'; the energy of the M_I state is therefore less than that of the L_{II} state by the amount $h\nu'$; calculating it thus, and then applying to normal atoms a stream of electrons or of quanta having energy just adequate to put them over into this M_I state, we find that this effect is duly produced.

Thus there is a thoroughgoing analogy between the genesis of optical spectra by transitions between the optical Stationary States, and the genesis of X-ray spectra by transitions between the X-ray Stationary States. The differences between the two kinds of spectra seem all to derive from the one fundamental difference between the two kinds of Stationary States; the former do not involve the absence of an electron from an atom, the latter do. In the optical region, for instance, we find that an atom in the normal state cannot be put into a particular excited state by any radiation except one of just the right frequency ν_o for which $h\nu_o$ is equal to the energy-difference between the normal state and the excited state in question. In the X-ray region, we find that an atom can be put into the K-state (for instance) by any radiation of frequency equal to or exceeding that critical frequency ν_o for which $h\nu_o$ is equal to the energy-difference between the normal state and the K state. This difference in behavior occurs because in the former case a quantum of radiation having frequency ν exceeding ν_o would have no place to put the left-over energy $h(\nu-\nu_o)$, whereas in the latter case this extra energy can be and is delivered over to the extracted electron as kinetic

energy, with which it flies away. This is known positively; for the extracted electrons can be observed, and their energy measured.

Spontaneous transitions from each of the X-ray Stationary States occur to some, but not to all, of the States of lesser energy. Some are evidently inhibited; and it is possible to lay down rules of selection, distinguishing those which are permitted. The complicated system of rules originally proposed has yielded place to a much simpler one, exactly similar to the one prevailing in the optical spectra. That is to say: it appears to be possible to assign to each of the X-ray Stationary States a certain value of a numeral k and a certain value of a numeral j, such that the only transitions which actually occur are those in which k changes by one unit and j either changes by one unit or does not change at all; while transitions between states in both of which $j=0$ are specially excluded. Furthermore, the various values of k and j thus assigned to the several X-ray Stationary States are identical with those assigned to the several States constituting a doublet system, such as we have met already in Section S, such as the sodium atom possesses; so that there is a complete correspondence between the system of X-ray Stationary States which every atom rich in electrons possesses, and the doublet system of optical Stationary States which only certain atoms possess. A part of this correspondence is expressed in the following Table:

TABLE II

Values of k:	1	1	2	2	1	2	2	3	3
Values of j:	1	1	1	2	1	1	2	2	3
Stationary States of									
Doublet system:	1s	2s	$2p_1$	$2p_2$	3s	$3p_1$	$3p_2$	$3d_1$	$3d_2$
X-ray system:	K	L_I	L_{II}	L_{III}	M_I	M_{II}	M_{III}	M_{IV}	M_V

No doubt the implications of this close correspondence are deep; but just what they are is not yet obvious.

The fact that the residue, left behind when an electron is extracted from an atom, may exist in any one of several distinct States, is quite naturally interpreted as meaning that the various electrons of the complete atom are variously situated, or revolving in various distinct orbits; so that the several X-ray Stationary States differ essentially in this, that differently-located electrons have been removed, leaving different places untenanted. This notion is easily combined with the idea that an atom is formed, or at all events behaves as though it had been formed, by successive self-annexations of electrons to a nucleus originally bare. Suppose that an atom is made by con-

secutive adhesions of electrons, each of which settles down into a peculiar orbit and remains there more or less unperturbed as the later comers immigrate one after the other into the system. May not then the process of X-ray absorption consist in a powerful intruding entity, electron or quantum, violently invading the interior regions of the atom and casting out one or another of the deeper-lying earlier-added electrons, while the later-added ones nearer or upon the frontier remain attached? May not X-ray emission consist in the passage of one of these latter electrons into the orbit formerly held by its predecessor, now unexpectedly reft away and its place left empty?

Although an affirmative answer to these questions involves a very literal and concrete conception of electron-orbits, most physicists make it, and would like to prove it. The chief difficulty lies in the fact that all information about X-ray Stationary States is primarily information, not about the prior condition of the electron which is gone, but about the final condition of the residue which is left behind.

The data show at all events that there are not nearly so many conditions of the residue, as there are electrons of the completed atom; from which it is fairly safe to conclude that the electrons of the atom are so arranged, that any one of several different electrons may be removed and the residue be left always in the same condition—therefore, that the electrons are arranged in groups, each electron being situated essentially like every other of its group. In discussing the formation of an atom by successive binding of electrons, it was remarked that several electrons may be bound in orbits each characterized by a common value of n and a common value of k. These two ideas may coincide; and great efforts are being made to bring them into entire coincidence. The evidence indicates, for example, that the first ten electrons bound to a nucleus are divided into four groups. Absence of an electron from one of these groups entails that the atom is in the K state; absence of an electron from the second, third, or fourth group brings it about that the atom is in the L_I, or L_{II}, or L_{III} state, respectively. So much the X-ray data do show rather definitely; although the actual number of electrons in each of the four groups cannot yet be deduced. If one could prove *a priori* that the first ten electrons annexed by a nucleus settle down into orbits of four distinct kinds, the achievement would be a great one. Intimations that something of this sort has been achieved are made every now and then; but it is difficult to tell whether the assertions which are made have been derived cogently from a principle or are inspired guesswork.

There is a remarkable numerical agreement in this field, the meaning

of which was until a couple of years ago regarded as perfectly distinct; but at this moment it is beclouded by one of the curious contradictions so abundant in the Theory of Atomic Structure. Briefly, the typical phenomenon is this: the differences between the energy-values of the K, L_{II} and L_{III} states agree notably well with what would be expected *if* the complete atom contains a few electrons moving in 1_1 orbits, a few in 2_1 orbits and a few in 2_2 orbits about the nucleus; and *if* the K state corresponds to absence of an electron of the first group, the L_{II} state to absence of one out of the second and the L_{III} state to absence of one out of the third. (The reason why calculations can be made for so indefinitely-phrased a model is this, that the field due to the highly charged nucleus of a massive atom should dominate over those of the individual electrons so that it does not make very much difference how many are supposed to be in each group.) The natural inference is, that the rest of the atom remains unchanged or little changed when any one of these electrons is extracted. In this case the Azimuthal Quantum-number of the residue should differ by one unit for the two states L_{II} and L_{III}. Consulting Table II, one finds that the quantity called k, which obeys the characteristic selection rule of the Azimuthal Quantum-number, is the same for L_{II} as for L_{III}. This is an illustration of the collisions between two sets of inferences which unsettle the supposedly firmest achievements of this theory.

Of the *theory of molecules*, a subject large enough for an article by itself, I can say here nothing more than that it attains some remarkable successes, achieved by and therefore fortifying some of the assumptions made in these pages; notably the assumption that Angular Momentum is a thing required in Nature to assume discrete values spaced at intervals of $h/2\pi$.

The final part of this long article has been very unlike the Second Part, in which an atom-model for the atoms of hydrogen and ionized-helium was constructed and endowed with certain fundamental qualities, so that it reproduced almost all of the relations of these atoms to radiation with a truly striking fidelity. This Third Part by contrast has been a thing of shreds and patches. Models for many atoms have been brought forth, but they have not been thoroughly adequate and they have not been concordant with one another. Some were designed with the same fundamental qualities as those given to the model for hydrogen; and scarcely more can be said for any of them, than that it does not positively clash with the properties of the element for which it is devised. Others were made competent to deal with a

certain limited set of facts (as the grouping of terms in multiplets) by giving them qualities gravely in discord with those attributed to the atom-model for hydrogen; and then they proved themselves surprisingly well able to account for isolated facts of quite a different sort (as the effect of magnetic fields upon atoms). The presentation in these pages is naturally very far from complete; had it been complete, it would have filled a book and not an article. But if it had been complete, the eventual impression would have been the same; an impression of confusion, yet of a confusion full of hope.

For the "Theory of Atomic Structure" is distinguished especially by this, that it is not one theory but a multitude of partial theories, each designed and competent to cover a limited family of the abounding data, each struggling to overlap and to absorb the others. It may be compared with a cross word-puzzle or a map-puzzle, in which the beginnings of a solution have been made in half-a-dozen corners and patches, while wide blank areas adjoin and separate them, and some of the partial solutions already entered upon the field may finally yield to others which can be unified into the perfected pattern. Or it may be compared with those maps of polar regions, in which here and there a properly-surveyed island or little strip of coastline emerges from the blankness of the unexplored realms, and some of them are certainly misplaced relatively to the others and will be shifted on the map when all the geography is at length made known. Or it may be compared with the state of a congealing metal, in which a multitude of little crystals have formed themselves about casual nuclei of crystallization; each is oriented in a different way, and when two of them grow into contact with one another they clash and cannot merge, they stand blocking and thwarting one another. It may be necessary to reliquefy them all and make a new attempt to change the formless mass into a single crystal.

Meanwhile the work is driven forward with the fervor of discovery and exploration, in this period which Russell finely called the "Heroic Age of Spectroscopy," and not of spectroscopy alone. Many, though not so many as are needed, are busy with determining the Stationary States by deciphering the rich and cryptic spectra of some among the numerous unstudied elements—enormously numerous, taking into account how many kinds of ionized atoms there are; and others with the assembling of new photographs of spectra made under the most varied sorts of excitation, with other aids to discriminating the lines; and others with the impressing of electric and magnetic fields upon radiating atoms; and others are engaged in measuring the intensities of lines. Yet other experimenters are determining the magnetic

moments of various atoms in all the possible ways. Some are seeking new phenomena which may result from Stationary States and from transitions, and occasionally they are rewarded with brilliant examples such as that vivid demonstration of the atom-magnets which Gerlach effected, or such as the passage of an atom from one State to another while it transfers the liberated energy to another particle directly and produces a chemical change. Others are finding the processes resulting from the Stationary States manifest on an unearthly scale within the stars.

The theorists likewise are at work with furious industry. Now and then a set of data hitherto rebellious is suddenly systematized, usually in a manner not quite concordant with the other theories holding other parts of the field. Attempts are made to unify one partial theory with another, usually unsuccessful. Sometimes an authoritative thinker, despondent over the continuing contradictions, tries to cut all the knots by declaring that one or another of the conflicting models is entirely fallacious, and that the numerical agreements on which it is founded are a delusion and a snare. Another is driven to concede that the conflicts are destined to endure forever, and accepts all of the partial theories as equally valid, or else paraphrases them in ingenious words which veil the contradictions, yet leaving these essentially unabated. Others, abandoning the general problem, have returned to the question of the hydrogen atom, and for this they are trying to rephrase or reshape the Quantum Conditions in a manner more satisfactory to themselves; sometimes with the aid of new and unfamiliar forms of mathematics, apparently expecting that when these become habitual to the human mind, the mystery of the Quantum Conditions will seem simple and clear. That, of course, always remains a possibility—that the human intellect will accustom itself so thoroughly to the new systems of ideas that they will cease to seem incoherent, as the human ear has so accustomed itself to the harmonic innovations of successive generations of musicians that the tones which seemed outrageous discords to the audiences of Beethoven now are to us monotonously sweet. To our minds the various divisions of the Atomic Theory are still discordant. It would not be fair to leave any other impression of this strange and fascinating theory; inchoate but full of promise, immature but gathering force, a fantastic assemblage of failures and successes; irreconcilable with all other theories, irreconcilable even with itself, and yet perhaps predestined to refashion all the science of physics in its own image.

Some Studies in Radio Broadcast Transmission[1]

By RALPH BOWN, DeLOSS K. MARTIN
and RALPH K. POTTER

Synopsis: The paper is based on radio transmission tests from station 2XB in New York City to two outlying field stations. It is a detailed study of fading and distortion of radio signals under night time conditions in a particular region which may or may not be typical.

Night time fading tests using constant single frequencies and bands of frequencies in which the receiving observations were recorded by oscillograph show that the fading is selective. By selective fading it is meant that different frequencies do not fade together. From the regularity of the frequency relation between the frequencies which fade together it is concluded that the selective fading is caused by wave interference. The signals appear to reach the receiving point by at least two paths of different lengths. The paths change slowly with reference to each other so that at different times the component waves add or neutralize, going through these conditions progressively. The two major paths by which the interfering waves travel are calculated to have a difference in length of the order of 135 kilometers for the conditions of the tests. Since this difference is greater than the distance directly from transmitter to receiver it is assumed that one path at least must follow a circuitous route, probably reaching upward through higher atmospheric regions. Various theories to explain this are briefly reviewed.

The territory about one of the receiving test stations in Connecticut is found under day time conditions to be the seat of a gigantic fixed wave interference or diffraction pattern caused in part by the shadowing of a group of high buildings in New York City. The influence of this pattern on night time fading is discussed. It is considered a contributing but not the controlling effect.

Tests using transmission from an ordinary type of broadcasting transmitter show that such transmitters have a dynamic frequency instability or frequency modulation combined with the amplitude modulation. At night the wave interference effects which produce selective fading result in distortion of the signals when frequency modulation is present. It is shown that stabilizing the transmitter frequency eliminates this distortion. A theory explaining the action is given. The distortions predicted by the theory check with the actual distortions observed.

A discussion of ordinary modulated carrier transmission, carrier suppression, and single side band transmission is given in relation to selective fading It is shown that the use of a carrier suppression system should reduce fading.

ONE of the factors which must be given increasing attention, if the technique of radio telephone broadcasting is to consolidate and continue its remarkable progress, is the mechanism of the transmission of radio signals through space. In many receiving situations the largest apparent defects present in the reproduced signal are those suffered not in the terminal apparatus but in transit through space, and in these cases better methods of utilizing the transmitting medium must precede any major betterment in overall results. In the present paper we are reporting some investigations in this field of radio transmission which have uncovered a number of interesting facts and have led to at least one conclusion which is of practical utility.

[1] Presented before the Institute of Radio Engineers, New York, Nov. 4, 1925

Night time transmission, which is the usual case in broadcasting, is in many places commonly marred by fading and sometimes by actual distortion of signals. Often these occur in certain areas not more distant from the transmitting station than other areas which enjoy freedom from such annoyance. Selecting a particular instance of these difficulties in an area near New York City which, in so far as can be judged at present, is probably a typical instance, we have subjected it to an intensive experimental study to determine what is the inherent nature of the troubles and if possible how they may be alleviated. In doing this it has been necessary to employ novel forms of tests especially fitted to bring out in a concrete way the phenomena being investigated.

To provide a suitable background for the subject we have started our discussion below with a brief recital of some of the things which a transmission medium is called upon to do. Following this we have described our tests, pointing out in what ways the existing media seem to fall short of doing these things and offering certain speculations as to the reasons for the shortcomings. In conclusion we have analyzed some practical problems in the light of this work.

Fundamental Considerations

As the radio art has progressed from spark telegraphy into continuous wave telegraphy and into high quality radio telephone broadcasting, increasing demands have been made on the transmission medium to deliver at the receiving point a true sample of what was put into it at the transmitting station. The requirements have grown in rigor because in telegraphy the end has been to develop increased reliability of communication at longer ranges and in telephony the medium is called upon to transmit a highly complex form of intelligence.

Of the requirements placed on the transmission medium by modern uses, those imposed by telephony are far more exacting than those for telegraphy. In telegraphy a single frequency, or at most a narrow band of frequencies sent out intermittently in accordance with a dot and dash code must reach the receiving station in such shape that it may be converted into audible sound for aural interpretation or into current pulses for the operation of relays or recording instruments. Leaving aside noise, the principal requirement is a sufficient freedom from fading so that signals can be interpreted or recorded without interruption. In radio telephony, as at present practiced in broadcasting, there is transmitted a modulated high-frequency wave comprising a relatively wide band of frequencies, usually at least 10 kilocycles.

Such a modulated high-frequency wave drawn out in the familiar graphical representation is a comparatively simple-looking thing, but analyzed into its elements and studied in detail it is revealed as being an intricate fabric of elemental waves so interwoven with each other that no one of them can be disturbed without changing in some degree the complexion of the whole. For perfect results the whole band must arrive at the receiver with an amplitude continuously proportional to that leaving the transmitter, or the inflections or expression of the speech or music will not be correctly reproduced. All the component frequencies within the band must be unchanged in their relative amplitudes lest the character of the sounds be altered. Even the relative phase relations of the various frequencies must be preserved or, as will be shown later, the interaction of the two side bands in the receiving detector will result in the partial loss of some of the frequency components.

It is not long since the time when radio was supposed to be the perfect medium for voice transmission it being presumed that since the ether of space (if there be such a thing) was substantially perfect in its electrical characteristics it must transmit frequency bands carrying telephone channels without distortion of any kind. This may be true theoretically of a pure ether but in fact, the ether used for radio communication is filled with a number of things ranging from gaseous ions down to the solid bed rock of the earth. It is rather to be expected that these will affect the progress of electromagnetic waves and we know from experience that they do. Diurnal variations of attenuation, fading, directional changes, dead spots and the like are already well known phenomena resulting from the complexity of our transmission media, although no entirely adequate explanations of their causes have been certainly established. One of the most recent manifestations of the effects of irregularities in transmission through space is in the distortion of the quality of telephone signals. This was perhaps first noticed in the use of short waves for broadcasting it being found that frequently the transmission was so distorted that after detection the signals such as speech and music were in severe cases almost unrecognizable.

Preliminary Investigations

For some time after quality distortion was recognized as a characteristic of existing short wave transmissions, it was thought that for the lower broadcasting frequencies at least, it was present only at night and at relatively very great distances from the transmitter. However,

careful observations demonstrated that there were points relatively near New York City where quality distortion from several broadcasting stations in the city was marked at night and in at least one case was detectable even in daytime. When station 2XB the Bell Telephone Laboratories' experimental station at 463 West Street, New York City, was used to transmit test signals, it was found that quality distortion could be observed in northern Westchester county and in southern Connecticut at distances of about 30 to 50 miles from the transmitter. Fading was also pronounced and it was noted as a significant fact that distortion was always accompanied by some fading although the reverse was not consistently true. In the course of these trials it was noticed that at a particular point near New Canaan, Connecticut, signals from 2XB were much weaker and more distorted than signals from 2XY, the experimental station of the American Telephone and Telegraph Company at 24 Walker Street, New York, even though the transmitter at 2XB was about ten times more powerful. Daylight field strength measurements at this point showed that the field strength of 2XB was only one-third that of 2XY. This led to the rather startling conclusion that there is a ratio of 100 to 1 in the power efficiency of transmission to that particular receiving point from these two transmitting stations in New York which are only about one mile apart.

In order to throw some light on this state of affairs a field strength survey was made by G. D. Gillett which resulted in the field strength contour map [1] here reproduced in Fig. 1. The contours on this map show that there is a series of long nearly parallel hills and valleys of field strength which, extrapolated, would converge in lower Manhattan and which extend out to the northeast as far as it was thought worth while to follow them. There has occurred to us no better explanation of this hitherto uncharted form of field strength distribution than that it is a gigantic wave interference pattern. A detailed discussion of this theory is given in another section of this paper.

The fixed pattern shown by Fig. 1 is definitely present only in the daytime but that it is fixed is attested by the fact that a second survey made about a year later checks with the original one quite closely. At night fading is pronounced in the area covered by the pattern and it is apparent that some other factors must enter. As a result of an endeavor to check up the pattern at night it was discovered that

[1] This map was prepared by Mr. Gillett using the methods discussed in a paper "Distribution of Radio Waves from Broadcasting Stations Over City Districts," by Ralph Bown and G. D. Gillett, I. R. E. Proc., Vol. 12, No. 4, p. 395—August, 1924.

quality distortion was, in general, most evident at places which were, by day, in the valleys of the field strength diagram and a point in one of these valleys near Stamford, Connecticut, was selected for the establishment of a temporary field test station. The interior of this station, which was in the empty hay-mow of a barn, is illustrated by the

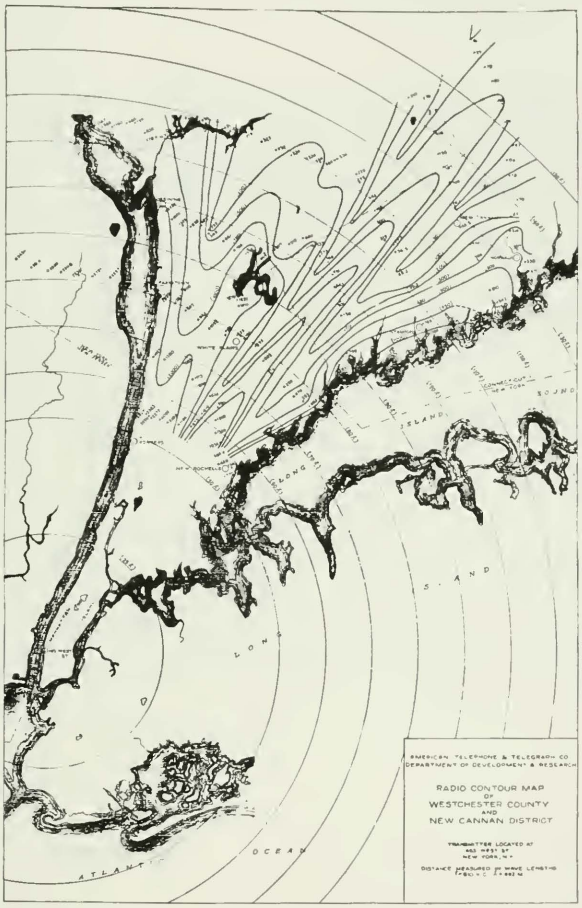


Fig. 1—Radio contour map showing wave interference pattern

photograph, Fig. 2. At this place apparatus was set up to enable a study of the nature of the distortion in signals from 2XB. Many of the records discussed in succeeding paragraphs were taken at this Stamford field station. Others were taken near Riverhead, Long Island, which was also found to be well located for such work. Fig. 3

is an outline map showing the relative positions of these field receiving stations and the transmitting station.

The reason for settling down at a fixed point in this way was to attack the problem from a new angle. The field strength survey and aural observations had yielded much interesting information but did not appear at that time to shed a great deal of light on the quality distortion so it was decided to attempt, by an oscillographic

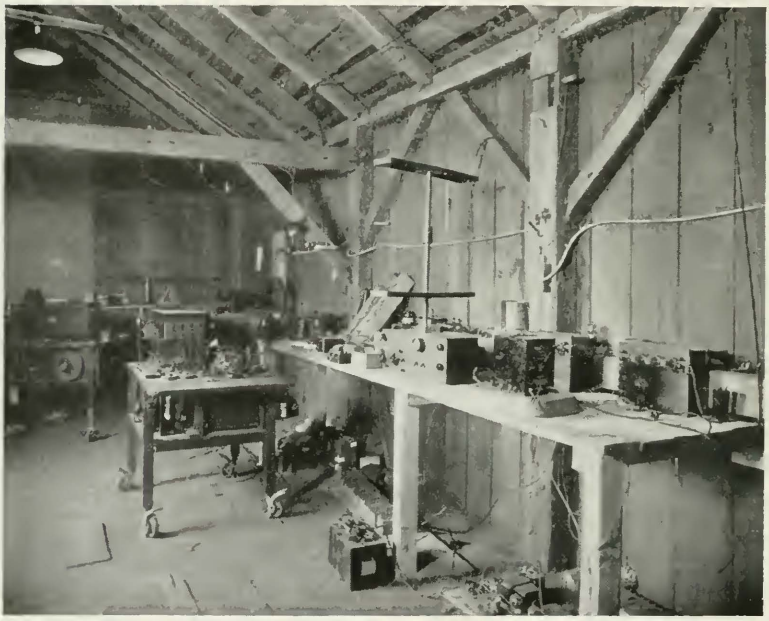


Fig. 2—Interior view of test station near Stamford, Conn.

study of received signals sent out under rigorously controlled conditions, to determine just what alterations these signals suffered in their journey through space.

In finding such distortions the ear is, of course, the primary testing instrument or indicator of trouble, for, if the trained ear is unable to detect anything wrong with a received signal in comparison with its original counterpart it is safe to say that nothing detrimental of importance has happened to it. But the ear is a poor quantitative indicator and furnishes no permanent or easily analyzed record of its observations. It is evident that if we are to study quantitatively the characteristics of radio transmission which give rise to quality distortion,

we must devise tests which will disclose changes, of whatever kind, in the relations between the various component frequencies of the transmitted band and furnish interpretable permanent records. In



Fig. 3—Outline map showing locations of transmitting station and receiving test stations

fact in the studies described herein a considerable portion of the job was to devise or perfect suitable methods of attack.

Single, Double, and Triple Frequency Tests

The variable factors in radio transmission which may be directly controlled are located at the transmitter and receiver. We have as yet no tangible means of controlling the transmitting medium, but it can be studied indirectly through the characteristics of the received signals. Obviously, it is desirable in the interest of simplicity to stabilize the apparatus variables to the extent that they may be idealized in considering observed results. Furthermore, at both the transmitter and receiver, it is desirable to make the antenna arrangements of the simplest form. For our work the normal antenna arrangement at station 2XB was used perforce since any important changes would have constituted a major operation. It is far from a simple arrangement, as shown in Fig. 4 which is an outline elevation and plan of the antenna and building at 463 West Street, New York City. Fortunately there are no buildings considerably higher than the antenna within a distance of several wave lengths.

At the receiving test stations both loop and vertical antenna were used; but in most of the experiments a simple vertical antenna was employed. It was constructed of brass tubing, 30 feet long, and guyed in a vertical position. A galvanized iron pipe 12 feet long was driven in the earth for a ground connection. The vertical receiving antenna projected through the roof of the test station building

at Riverhead, L. I., as shown in Fig. 5. The receiving antenna was not tuned but was connected to the radio receiver through fixed inductive coupling.

The carrier power in the transmitting antenna normally remains fairly constant, except for minor variations in voltage of the supply mains, and with a little care on the part of operating personnel, the

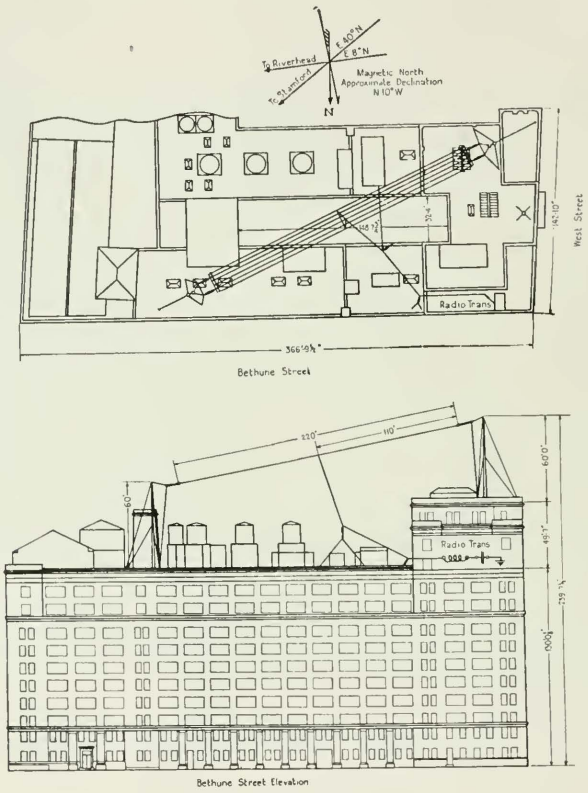


Fig. 4—Plan—elevation of the transmitting antenna

antenna current can be kept within the limits of a 1 per cent. variation, which is small compared with the signal fading usually experienced.

The stabilization of the frequency was of the greatest importance since in some of the tests it was desired to beat or heterodyne the signals down to audio frequencies and pass them through narrow band filters. To provide this stability engineers of the Bell Telephone Laboratories arranged the 5-kw. transmitter at station 2XB to obtain

its carrier frequency by amplification of the output of a 610-kc. piezo-electric crystal oscillator.

When desired some of the antenna current from the output of the transmitter was rectified and the resulting current was sent over a telephone line to the receiving station so that the frequency and wave



Fig. 5—Receiving test station near Riverhead, L. I. showing vertical antenna projecting through roof of building

form of the modulating signal could be seen and photographed at that point, thus guarding against any possible distortion in the transmitter and enabling a direct "before and after" comparison to be made. The telephone circuit was also used for communication between engineers at the two terminal stations.

At the receiving station double detection receivers and audio frequency amplifiers were employed. These did not have entirely "flat" transmission characteristics over the audio frequency band, but in most of the tests this was of no importance. In cases where it affected the results the making of necessary corrections was a simple matter. In tests involving beating the received signals down to audio frequencies through the agency of a local heterodyning frequency,

this was supplied from a shielded vacuum tube oscillator which on comparison with a standardized piezo-electric oscillator was found to possess the required stability. The double detection type receivers were used for no other reasons than their availability and their convenience for quantitative work. The beating down oscillator within the sets and the intermediate frequency step passed through in the sets by received signals do not figure in the following discussion of test methods but, of course, in each case the necessary set tuning

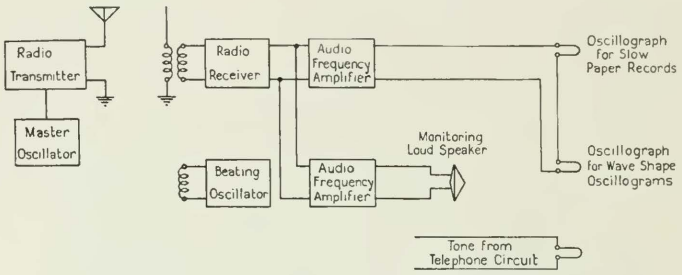


Fig. 6—Diagram of system used for single frequency tests

adjustments were made. To avoid confusion it is well to think of these receivers as being replaced by high frequency amplifiers and simple detectors since the local beating oscillator referred to in later pages is the separate shielded oscillator described above which is used to beat the signals down to audio frequencies.

In this work the moving coil type oscillograph was used throughout for the purpose of making photographic signal records. As indicated in Fig. 6 two oscillographs with elements connected in series were employed; one for the purpose of making a continuous record of the variation in the amplitude of the signal using a slow moving photographic paper tape and the other to obtain the wave shape of the signal by means of the usual high speed photographic film drum. An element of one oscillograph was also used at times to record on the film drum the wave shape of signals rectified at the transmitting antenna and sent over the telephone lines.

Fig. 7 is the interior view of the test station at Riverhead showing the general arrangement of the oscillographs and accessory apparatus. This oscillograph equipment formed about the only fixed portion of the apparatus, other portions being changed from time to time for different tests. These arrangements will be described later in connection with the records which they were used to obtain.

In considering these various records perhaps we had best look first at the simpler ones and then proceed in a more or less orderly

fashion to the more involved ones. The simplest records are fading records of the unmodulated carrier frequency of 610 kc. At the receiver the carrier was heterodyned with a local oscillator to produce a beat tone of about 250 cycles which was fed through amplifiers to the oscillograph elements.

A representative sample of the form of signal records made in the manner described above which show the variation in the amplitude of

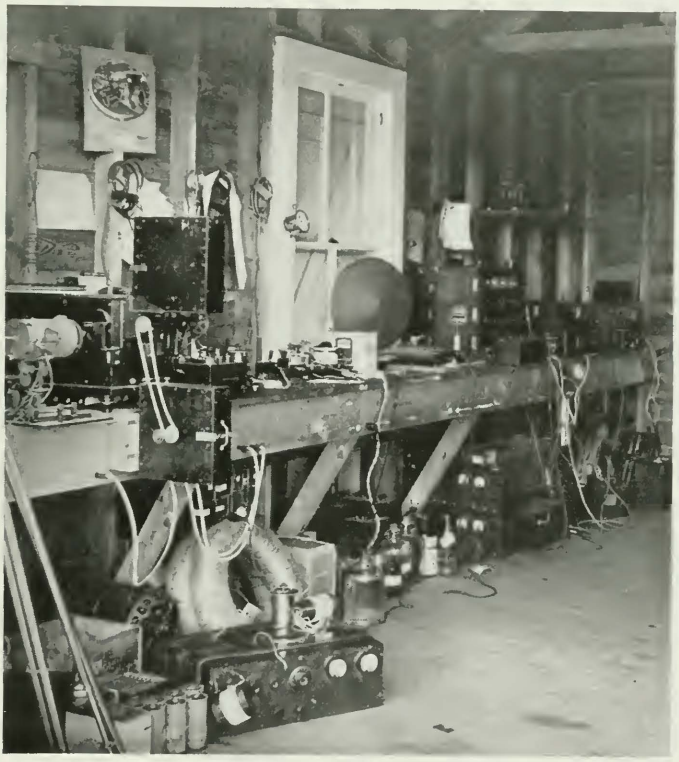


Fig. 7—Interior view of Riverhead testing station showing recording apparatus

the received carrier signal with time, is given in Fig. 8. It shows a typical fading record made at Stamford, Conn., May 16, 1925. The timing interval on strip 6 is 2.6 seconds.

The feed of the photographic paper tape through the oscillograph was varied somewhat during the course of the experiments but was generally in the range of 6 to 12 inches a minute. At this rate the record of an audible frequency signal is a shadow band of varying

width corresponding to twice the amplitude of the signal, as both the positive and negative half-cycles are recorded. It will be observed that the outer limits of the band corresponding to the peaks of the sine wave are darker than the center portion of the record. This is due to the fact that the rate of change of the movement of the light

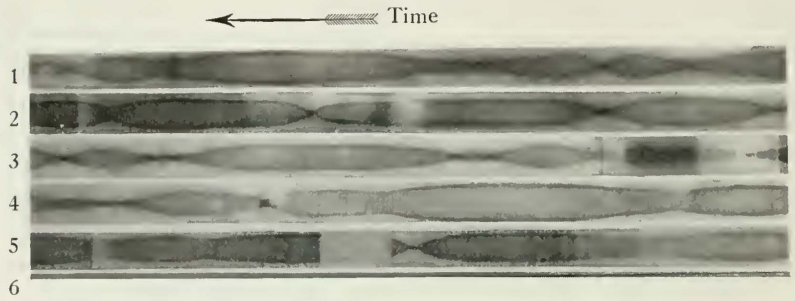


Fig. 8—Single-frequency fading record. Made at Stamford, Conn., May 16, 1924, 1:54 a.m. Timing marks, on strip 6, 2.5 seconds apart

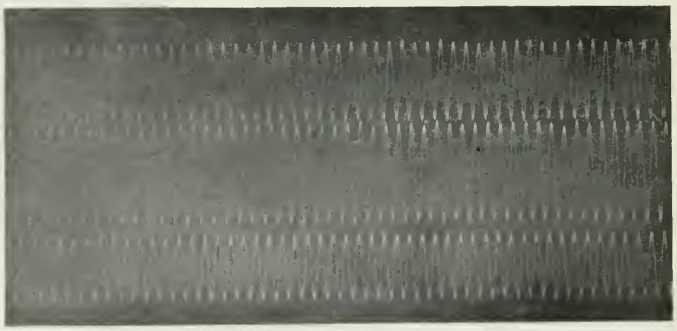


Fig. 9—Wave form of beat note signal for single-frequency test. Center trace signal from vertical antenna, upper and lower traces signals from loop antenna receivers

spot on the record is a minimum at the peak of the signal; hence, a greater quantity of light affects these portions of the record. This shading effect was very useful in the way it brought out changes in the distortion of the received signal. This is discussed fully in another section of the paper. The fuzzy irregular outline on portions of the records is caused by static and radio noise. The timing marks on the record allow a measurement of the time interval between points of minimum signal. Fig. 9 is a sample oscillogram of the wave shape of a beat note signal recorded by the method described above.

Marked changes in the fading cycle or time interval between points of minimum signal may occur within a period of a few minutes, and

from day to day there is often evidenced a modification of the general character and the recurrence of these changes. An example of this change in a short period of time is well illustrated by the oscillograms in Fig. 10. Strips 1, 2 and 3 form a continuous record starting at 1:52 a.m.; strips 4, 5 and 6 start at 2:16 a.m.; and strips 7, 8 and 9 start at 2:37 a.m. These are three sections of a continuous record

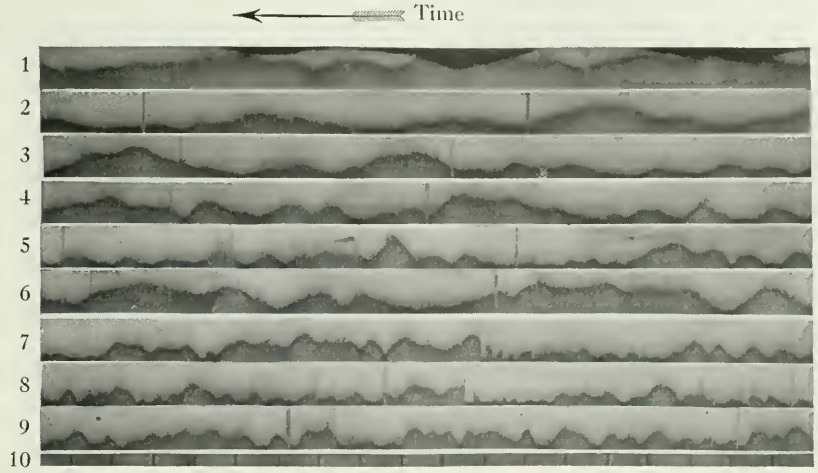


Fig. 10—Single-frequency fading record, showing variation in rapidity of fading, made at Riverhead, L. I., July 16, 1925, 1:52 a.m. Timing marks, on strip 10, 5 seconds apart

selected for the purpose of showing the decrease in the fading period, in a 45-minute interval. The timing interval on strip 10 which applies to these records is 5 seconds. In this particular record only half of the audio signal was recorded, the edge of the strip being the zero line.

These single frequency fading records do not offer very much to work on. There is, however, just enough suggestion of regularity about them to annoy one with the thought that perhaps they may follow some definite combination of periodicities and with this in mind we have taken sections of two different records and subjected them to a harmonic analysis.

So far we have been able to draw no more useful conclusions from such harmonic analyses than that the heterogeneous scattering of harmonic values is about what one would expect from the looks of the curves.

One significant thing about these oscillographic single frequency fading records is that they show no high speed fading of important

magnitudes. Occasionally one cycle of the beat tone will be somewhat upset by a sudden change in the amplitude but in general no changes which consistently distort the wave form were observed.

The slow fading may be considered as a modulation and on this basis the received signal is seen to be composed of the original constant carrier frequency accompanied by very narrow side bands occupying at best perhaps a fraction of a cycle.

The next progressive step in the radio transmission studies is naturally from a single frequency to two or more frequencies transmitted simultaneously. By the use of two crystal oscillators at the transmitter two separate and distinct radio frequency signals were transmitted simultaneously. These crystals were ground by the Bell Telephone Laboratories to oscillate at 610,000 cycles and 609,750 cycles. The amplitudes of these signals at the transmitter were controllable so that it was possible to make them equal, or one larger than the other, equivalent to the relative magnitudes usually found for the carrier and single side-band transmission case. Records were obtained of the variation of these radio signals, but none is reproduced here since the information shown by them can be just as easily obtained from the triple frequency records shown below.

Radio transmission on three frequencies is readily obtained by modulating the carrier with an audio frequency tone, and observing the three frequencies separately at the receiver.

If the modulating tone is

$$\sin (vt+\phi)$$

and the carrier signal

$$A \sin pt,$$

the transmitted signals are

$$+\frac{Aa}{2} \cos [(p+v)t+\phi] \qquad \text{(upper side band)}$$

$$+A \sin pt \qquad \text{(carrier)}$$

and

$$-\frac{Aa}{2} \cos [(p-v)t-\phi] \qquad \text{(lower side band).}$$

where a is a constant proportional to the percentage modulation.

These three frequencies are not merely a mathematical fiction but are physically existent as three separate waves bound together only at their point of origin.

To adequately record them separately by means of the oscillograph advantage was taken of the fact that a group of frequencies beaten

with a single frequency differing from them by a small amount and detected may thereby be reduced to audible frequencies without having their interrelations of phase, amplitude or difference frequency composition, changed in any respect. For instance if the frequencies expressed above are beaten with a local constant frequency,

$$B \cos (qt+\psi)$$

the resultant lower or difference frequencies will be

$$+\frac{kBAa}{2} \cos [(p+v-q)t+\phi-\psi]$$

$$+kBA \quad \sin [(p-q)t-\psi]$$

$$-\frac{kBAa}{2} \cos [(p-v-q)t-\phi-\psi].$$

Each one of the three components has been changed in amplitude by the same factor kB representing the efficiency of detection. Each

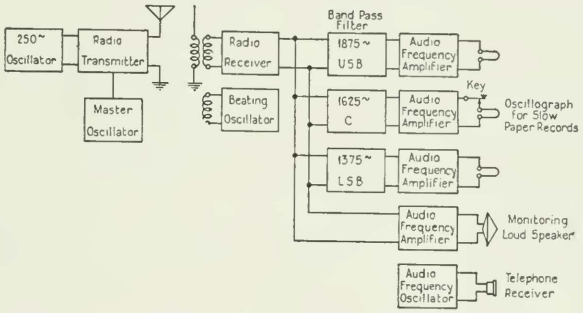


Fig. 11—Diagram of system used for three-frequency tests

one has been reduced in frequency by exactly the same amount $\frac{q}{2\pi}$ and each has had its instantaneous phase shifted by an angle $-\psi$. Relative to each other they remain unchanged.

In our actual case the carrier frequency $\frac{p}{2\pi}$ was 610 kc. The modulating frequency $\frac{v}{2\pi}$ was 250 cycles and the beating frequency $\frac{q}{2\pi}$ was 608,375 kc. so that the resulting three audio frequencies were 1,875 cycles, 1,625 cycles and 1,375 cycles.

As indicated in Fig. 11 in order to make a record of these signals they are separated at the receiver by means of band filters. These filters and others similar in type for other modulating frequencies

were designed and made by the Bell Telephone Laboratories especially for this work. The band filters used for the purpose of selecting the carrier and side-band frequencies had a cutoff of 40 Transmission Units 250 cycles from the mid-band frequency.

These cutoffs together with the position in the frequency range of the pass bands of the filters preclude any troubles from cross modulation of the radio carrier and side bands during the beating down process. The products of such cross modulation would be frequencies which are multiples of 250 cycles and these cannot pass the filters. On the other hand the beaten down frequencies will pass practically intact, since as has been shown by the previously described single frequency tests, each of the three frequencies received although subjected to amplitude modulation by fading, represents only a very narrow band of frequencies for which the filter pass bands were of adequate width.

As the modulating tone was carefully calibrated to 250 cycles and the filters adjusted to transmit the frequencies specified, it was only necessary to transmit the carrier while adjusting the receiving beating oscillator. The following procedure for this adjustment was found to be very successful. A local audio frequency oscillator was set to the reduced carrier frequency of 1,625 cycles, and its output connected to a telephone receiver. The audio beat note from the radio signal and local beating oscillator was reproduced by a loud speaker and its frequency adjusted to zero beat the 1,625-cycle tone from the telephone receiver.

When this adjustment had been completed the carrier was modulated with the 250-cycle tone, and the side-band signals automatically pass through their respective filters.

The signals from the outputs of the filters were amplified, and recorded separately by the three oscillograph elements. The sample records shown in Fig. 12 are representative.

Strips 1, 2 and 3 are taken from a long record obtained May 7, 1925, 3:22 a.m. The upper trace is a record of the upper side-band signal, the center trace the carrier, and the lower trace the lower side-band. Strips 4, 5 and 6 are from a section of a similar type of record made May 23, 1925, 1:06 a.m., where the carrier was modulated with a 500-cycle tone and different filters were used. In this record the upper trace is the lower side-band and the lower trace the upper side-band.

It will be noticed that the timing interruption appears only in the side-band signals, as the tone was interrupted before modulation took place, and that the amplitude of the carrier signal is not affected

by the interruption of the modulating tone. This makes it very easy to identify the side-band signals. These records give an excellent graphic picture of ordinary radio telephone transmission, bringing out the fact that three truly individual frequencies are transmitted to reproduce one.

In Fig. 12, strips 1, 2, and 3, the relative amplitudes of the three signals are very nearly in proportion to the relative amplitudes of

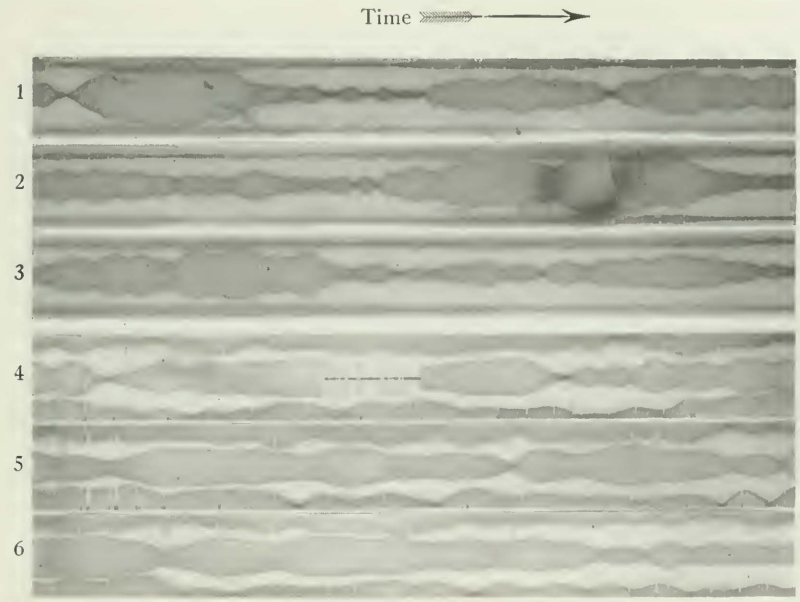


Fig. 12—Fading record showing individually the fading of carrier and side-band frequencies. Made at Riverhead, L. I. Timing interruptions in side-band signals, 5 seconds apart

the signals as they existed in the ether at the receiving point. Before this record was made a transmission characteristic of the complete receiving circuit, including the oscillograph elements, was obtained, using a local transmitter with modulated carrier for the purpose of making the measurement. The gain of the audio amplifiers at the outputs of the filters was adjusted to give substantially uniform transmission on each of the three frequencies corresponding to the carrier and side bands of the radio frequency signal.

As shown in Fig. 11, a telegraph key is placed in the circuit of the center oscillograph element, for the purpose of placing identifying signals on the records. An example of these identifying signals is

shown in Fig. 12, strip 4, which gives the date and time the record was started, July 23, 1925, 2:06 a.m. (Eastern daylight saving time).

The record in Fig. 13 is of the carrier and side-band signals with 500-cycle modulation made at Riverhead, L. I., May 25, 1925, 1:25 a.m. More gain was used in the side-band amplifiers for this record in order that the effects of fading could be brought out more prominently. In this record only half of the side-band signals were recorded, the

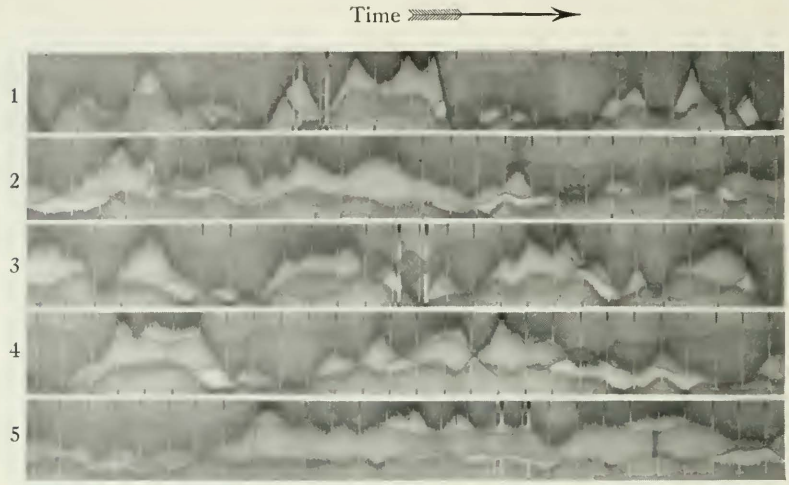


Fig. 13—Fading record of carrier and side-band signals, made at Riverhead, L. I. Timing interruptions in side-band signals, 5 seconds apart

zero reference line being at the edge of the strip. The upper trace is the upper side band, the center the carrier and lower trace the lower side band. Where the traces of the signals overlap a darker record is obtained. This record may be confusing at first but if strip 5 is examined where the amplitudes of the signals are not so large a better picture of the form of the record will be obtained.

It is obvious from these records that the carrier and side-band signals do not fade together as a unit. The carrier may pass through a zero value with still considerable amplitude in the side-band signals as in strips 1 and 3. In the first case, strip 1, the three frequencies successively fade through points of minimum signal in the order lower side-band, carrier and upper side-band; and in the second case, strip 3, the three frequencies fade through points of minimum signal in the reverse order. This is a definite indication of *selective fading*; that is, *fading is a function of frequency as well as time.*

An endeavor to form an explanation of the cause of this selective action in fading must be largely in the nature of speculation. Furthermore, since our data consist in the results of things which have happened rather than in any first hand information on the processes of the happening, the building of an explanation is a synthetic process. In general for any given set of facts it is possible to synthesize a number of explanations. Bearing this philosophy in mind we have considered various theories in connection with our observations and have concluded that simple wave interference as a major cause of the signal variations is at present the most likely explanation. While wave interference may be called a major cause it should perhaps also be called a secondary cause since the assumption of wave interference presupposes for its origin, primary causation by some physical state or configuration of the transmission medium. Speculation as to the nature of this primary cause is one stage further removed from the data contained in our oscillographic records than is the assumption of wave interference.

Since it is desirable in the remainder of this discussion to point out the evidences of wave interference, let us consider briefly the nature of this phenomenon.

To avoid any possible confusion of terms let it be said that by "wave interference" we mean a particular physical phenomenon in wave transmission and have no reference whatever to static, signals from other stations, or any other of the forms of radio noise which are commonly designated by the word "interference" when they hinder the reception of desired signals.

When two single frequency plane polarized wave trains start out at the same time from a common source and travel by different routes to meet again at a distant point the nature of disturbance at that point is determined by the relative space phases of the planes of polarization and time phases of the amplitude of the two arriving waves.

If we let E represent the vertical resultant of the electric field, which would be the only part affecting a simple vertical antenna, such as we have used in most of our tests, then

$$E = e_1 \sin 2\pi(Ft + d_1) + e_2 \sin 2\pi(Ft + d_2) \qquad (1)$$

where F is the frequency and d_1 and d_2 are the distances along the respective paths measured in wave lengths and e_1 and e_2 are the vertical components of the two waves. These two sine terms may be thought of as two vectors differing in phase.

The condition that these add giving a field

$$E=(e_1+e_2)\sin 2\pi Ft$$

$$\text{is that, } d_1-d_2=(\text{a whole number}) \tag{2}$$

that is, the difference in length of the two paths must be an exact whole number of wave lengths. The condition that the two waves cancel each other giving a field

$$E=(e_1-e_2)\sin 2\pi Ft$$

$$\text{is that, } d_1-d_2=(\text{a whole number})+\tfrac{1}{2} \tag{3}$$

that is, the difference in length of path must be an exact odd number of half wave lengths.

Thus if the two components e_1 and e_2 are equal, the resultant vertical field E will go through values ranging from (e_1+e_2) down to zero as the path lengths change relative to each other. If the two waves do not have exactly the same amplitude, the minimum value of E will be something more than zero.

Differences in attenuation of the two waves and differences in their direction of arrival will modify the relative amplitudes of e_1 and e_2 but will not modify the time relations required for minima of the resultant field E unless we assume that at the time of a minimum neither wave has an appreciable vertical component. Since the consequences of such an assumption do not accord with our experimental data we have considered that it may be left out of account in the present discussion.

This is obviously a picture which fits in very well with the simple single frequency fading records. The major maxima and minima occur when the conditions of equations (2) and (3) are met and e_1 and e_2 are nearly equal. On the other hand it seems doubtful that the picture can be so simple. If we suppose two wave paths why not three or more? Additional paths would add irregularities to the fading and it would not be necessary to assume as great a degree of irregularity in the changes in any one path. But with an increasing number of paths the various arriving waves would tend to average to a more or less constant mean value and large departures from this mean would become rare. The fact that the fading signal continually covers a large range of amplitude, with the maximum many times the minimum, definitely points toward there being but a very small number of major paths, probably not more than two.

Considering now the question of selective fading in relation to wave interference we refer back to equation (2).

If we assume the distances to be measured in any desired units and call them d_1' and d_2' our equation will still hold provided we divide each distance by the wave length measured in the same units, thus

$$\frac{d_1'-d_2'}{\lambda}=\text{a whole number}=x;$$

rearranging this and writing $\frac{V}{F}$ for λ where V equals the velocity of the waves, we have

$$F=x\left(\frac{V}{d_1'-d_2'}\right). \tag{4}$$

If now we assume $(d_1'-d_2')$ to be fixed we find that F can have a series of values which are integral multiples of $\frac{V}{d_1'-d_2'}$ which we may call the frequency spacing interval. That is, with changing frequency E will go through maximum values with frequency at a series of frequencies beginning theoretically with zero and extending upward in regular spacing to infinity.

The spacing interval is obviously that number of cycles which corresponds to the lowest finite frequency in the series, namely, the frequency for which the distance $(d_1'-d_2')$ is just one wave length since when $x=$unity equation (4) becomes

$$F_1=\frac{V}{d_1'-d_2'}=\text{the spacing interval.} \tag{5}$$

By using the same process on equation (3) we find that E has minimum or zero values at another series of frequencies having the same spacing interval but lying midway between the frequencies at which maxima occur.

Thus it is apparent that with fixed path length difference the amplitude of the field E will be different for different frequencies, ranging from maxima of (e_1+e_2) down to minima of zero if the polarization planes and amplitudes of the two vertical components are equal.

Furthermore, still thinking of equation (1) as representing two vectors, it is evident that the phase of the resultant field is different for different frequencies even though these different frequencies had exactly the same starting phase at the source.

If the paths are changing with time, the field at a given point, as has already been pointed out, will go through time fluctuations. Another way to look at this is that there is a space pattern of maxima

and minima and as the paths change the plane section of the pattern taken by the surface of the earth wanders so that at any one point the field is continually fading in and out as the maxima and minima glide by it. Each frequency has its own pattern differing from those of its neighboring frequencies in such a way that at any given point the relation between amplitude and frequency is that just discussed

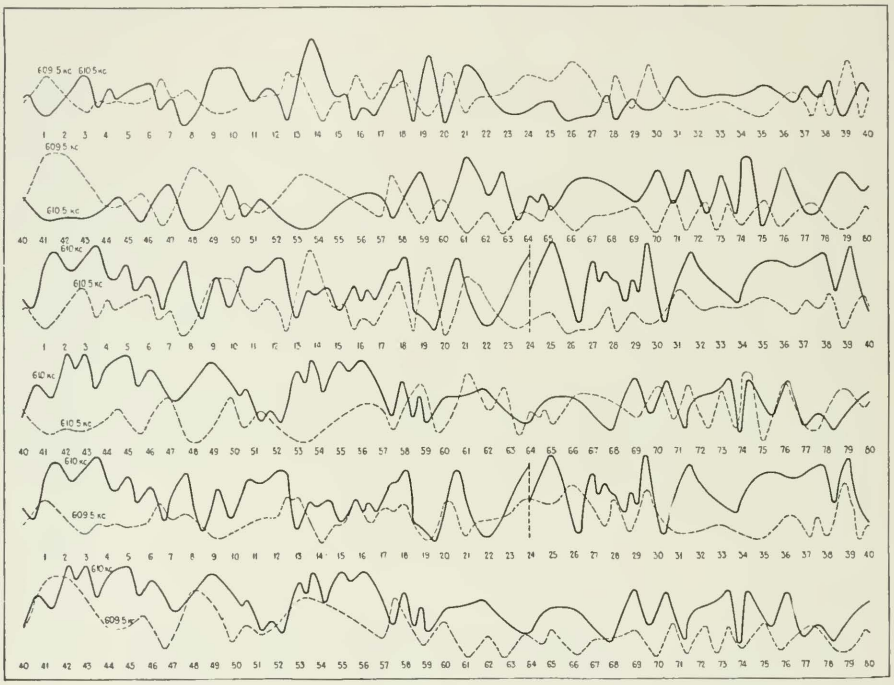


Fig. 14—Plotted curves of signal amplitudes condensing a long fading record, part of which is shown in Fig. 13. Numbers along time axis correspond to successive 25 second timing interruptions

above. Thus as the paths change and the patterns shift the different frequencies fade not simultaneously but progressively.

In the above analysis of wave interference it has been assumed that all frequencies traveled from transmitter to receiver over a given path in the same elapsed time. This does not mean that they necessarily follow exactly the same route on this path since they might follow somewhat different routes of equal length or if their transmission velocities were different they might follow different routes of unequal length and still come within the definition of a "path." It seems reasonable to assume that over the width of an

ordinary transmitted band the various frequencies are treated alike by the medium and the simple assumption that they follow the same route with the same velocity is justified. If none of these assumptions is correct but the departure is not large the effect will be merely to introduce slight irregularities into the spacing interval and the general nature of the result will not be changed.

Let us now examine more closely the record, a part of which is shown in Fig. 13. A portion of this has been condensed into the curves of Fig. 14. One unit along the time axes of these curves represents a 25-second interval.

To obtain these curves the amplitude of the signal has been scaled off and plotted, ignoring all the minor irregularities. From this record the relative fading characteristics of these single frequency signals 500 cycles apart are more easily seen, and it is possible to contrast the time of occurrence of points of minimum signal for any pair of them.

For the frequency difference of 500 cycles (610.5–610 and 610–609.5) these times are obviously quite different but there is no clearly

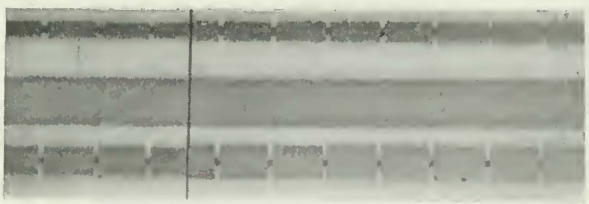


Fig. 15—Daytime record of carrier and side-band signals

discernible relation between them. The curves for 1000-cycle difference (609.5–610.5), however, show a striking relation in that the maxima and minima of the two are opposed fairly regularly over the entire 33-minute interval covered by the plot. This means that when one frequency has a minimum amplitude the other has a maximum and vice versa. Certainly this suggests a wave interference involving only two major paths whose difference in length is such that the spacing interval is 2,000 cycles. The path difference appears to be changing somewhat irregularly but at an average rate of the order of one wave length (or approximately 500 meters) per minute.

Before speculating further on the numerical values which may be derived from this part of the data we had perhaps best consider some other records of a somewhat different kind which are better adapted to provide such values. But first let us reiterate that these are *night-time* effects.

During the day signals substantially uniform in amplitude are received. An example of the type of transmission obtained in the daytime is given in Fig. 15, which is a record of the carrier and side-band signals received with substantially the same terminal conditions with the exception of the time as that existed when the records shown in Fig. 12 were made.

The abrupt change in the amplitude of the side-band signals was due to an intentional change at the transmitter in the input level of the tone modulating the carrier, and accordingly the amplitude of the carrier did not change. The timing interval is 5 seconds.

Band Fading Records

The familiar fading record is limited to two axes, amplitude and time. So far we have extended this cramped perspective somewhat by observing as many as three separate fading records spaced at audio-frequency intervals along the frequency axis. Even these three narrow lookouts upon the wide range of ether transmission have indicated amplitude relations along the frequency axis which

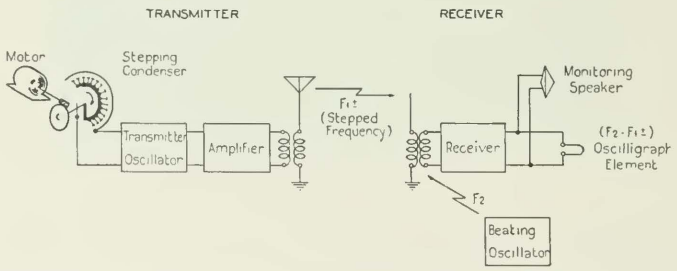


Fig. 16—Diagram of system used to obtain records of selective fading or "band fading" records

promise to open a new line of attack upon the problem of night-time fading. But the desirability of knowing what takes place in the interval unrevealed by these cracks in the fence becomes obvious. We should like to know the relative amplitude of frequencies over a wide band, and the change in this relation with time.

Since it is not a simple matter to record simultaneously the amplitude of a large number of waves of frequencies separated by say one hundred cycles in the radio-frequency range a single frequency in combination with a frequency stepping device at the transmitter has been adopted. The circuit arrangement is shown diagrammatically in Fig. 16. The rotary contactor bringing into the circuit suc-

cessively a total of fifteen small condensers across the main condenser of the transmitter oscillator shifts the frequency in steps over an adjustable range. The contactor is rotated at the rate of nine revolutions a minute, which is sufficiently slow to show definite steps in the oscillograph record. At the receiving end a local oscillator supplies a radio-frequency wave for beating the incoming frequencies down to values within the audible range.

A long oscillograph record of this stepped frequency gives a sort of moving picture of the fading for the entire band covered. A sample of such a record is shown in Fig. 17 with alternate pictures in the series removed to simplify the relations, since by reason of

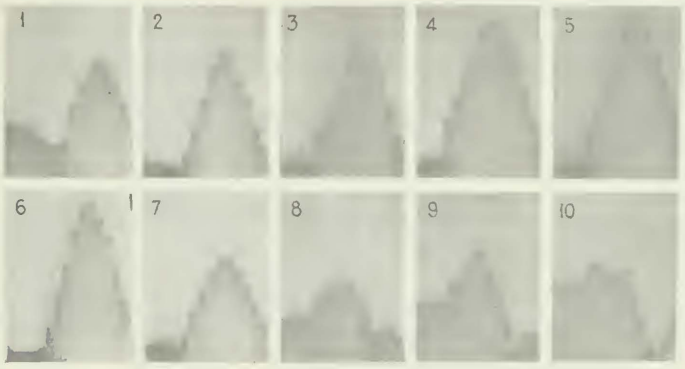


Fig. 17—Sample band fading record

the two-way traversal of the frequency band successive pictures are reversed. If a series of such built-up pictures as these could be taken rapidly on moving picture film, and projected successively upon a screen we should have before us an animated view of band fading. And according to the results of experimental investigation the subject offers a lively theme for such a presentation. The peaks and depressions glide nervously back and forth across the setting. The successive pictures of Fig. 17 (which, by the way, were selected for their half-tone reproduction possibilities rather than as first class examples of the records taken) illustrate a rather leisurely movement of this sort. These ten built-up photographs cover a period of slightly more than one minute. In the first seven pictures a depression appears at the left, while in the last three this depression seems to have made an exit followed by the simultaneous entrance of another from the opposite wing of the stage. Evidence of such

organized spacing of the minima is present in all of these night-time band fading records. As has already been suggested such evidence has an important significance, but before going into this phase of the subject again let us examine a little more in detail the structure of these band fading records.

The steps in any one picture of Fig. 17 are, as we have said, snap-shots of the wave amplitude for successively different radio frequencies taken about a quarter of a second apart. The fact that the fifteen snap-shots used to build up a single picture are not taken simultaneously causes a skewing of the outlines when movement of the depressions as shown in Fig. 17 occurs. If, for example, we

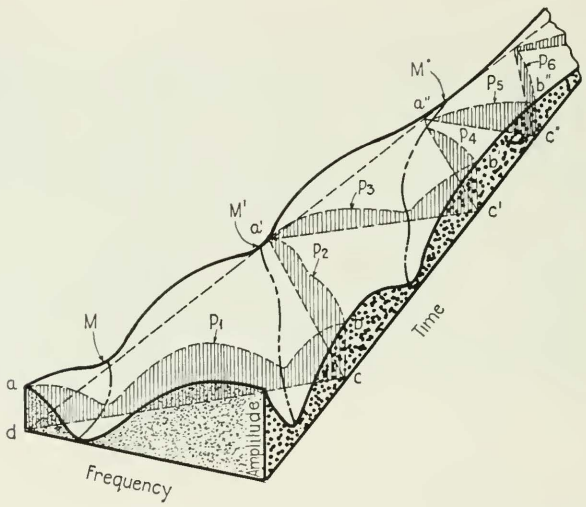


Fig. 18—Three dimensional diagram, showing the method of interpreting band fading records

were to take fifteen separate and successive snap-shots of a mountain through fifteen long vertical slits side by side it would be possible to combine the narrow sections so as to form a true picture of the peak. Now, if by some prodigious act of nature the mountain were shifted suddenly to one side and back again during the time we were taking the fifteen successive snap-shots through the vertical slits, the combination of them would form a profile quite different from that obtained when it was stationary. Or if it were simply moved steadily across the field of vision during the time the snap-shots were being taken one slope would be made to appear precipitous while the other would be leveled to a gentle grade in the finally built-up picture.

The character of this skewing, then, and its magnitude depend upon the rate at which the object being photographed in vertical sections moves, and the direction of the movement.

In Fig. 18 is shown an imaginary night-time band fading record in the "assembled" form. Since such a record contains frequency as a third dimension, in addition to amplitude and time as shown in the ordinary fading record, our simple fading curve has assumed the broader aspect of a surface, the selective fading making more or less parallel "valleys" running across it. The step-frequency system of recording the points amounts to photographing sections of this solid. The important point to be kept in mind is that these sections are *not perpendicular to the time axis.* If they were, the skewing previously described would not be present. By setting these sections up in their true relation to the time axis, however, and filling in to produce a continuous surface such as is shown in Fig. 18 the result is correctly represented. In order to make a detailed and accurate study of the band fading records, therefore, it is desirable to construct from the oscillograph sections the complete surface by the method suggested.

In Fig. 18 the trace of minima crossing the band is shown by M, M' and M''. Picture sections obtained as our recording apparatus literally moves back and forth across this frequency band are shown as $(a\text{–}b\text{–}c\text{–}d)$, $(b\text{–}c\text{–}a')$, $(a'\text{–}b'\text{–}c')$, etc. It will be evident that the section P_1, for example, will, in case a minimum is crossing rapidly, appear entirely unrelated to section P_2. When the minima run nearly parallel to the time axis (slow changes in transmission conditions) the successive pictures P_1, P_2, P_3, etc., will reveal their relation by direct comparison.

Actually to obtain frequency-amplitude sections perpendicular to the time axis in Fig. 18 would require the simultaneous transmission and reception of a large number of frequencies spaced at short intervals along the frequency axis. A more practical thought is to speed up the process and though this seems very simple at first consideration, it will be shown later to involve a particular kind of distortion which cannot be separated out as easily as the skewing encountered by the more deliberated method.

Now that we are familiar with the data, Fig. 19 showing, partially superimposed in vertical strips, the outlines of successive built-up pictures of the frequency traverse will be of greater significance. During the steady periods there appears within the 2,280-cycle band covered by these data approximately one complete cycle of selective fading. The lack of flatness in the audio-frequency-transmission characteristic of terminal apparatus has caused the suppression

of amplitudes toward the right side of these sections. Keeping in mind also the skewing inherent to this system of presentation during transient periods, we are able to trace the movement of minima, as illustrated previously in Fig. 17 which was taken from a different

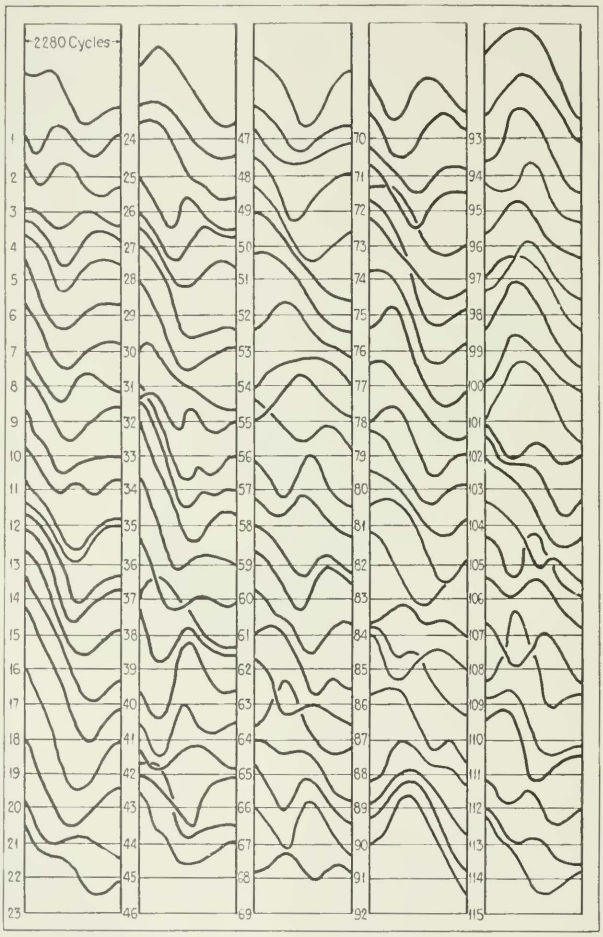


Fig. 19—Plotted curves, condensing a long band fading record so as to bring out the effect of selective fading

record. The relative position of these minima gives us an interesting insight into the nature of the night-time transmission path.

From records covering frequency ranges up to 4,500 cycles in width the positions of major minima along the frequency axis have

been plotted against time as in Fig. 20. The widths of the frequency bands covered in this case are indicated. This picture is essentially a bird's-eye view of band fading records such as are illustrated in idealized form by Fig. 18, the amplitude axis being perpendicular to the page. It reveals the presence of minima spaced at more or less definite frequency intervals, and suggests the presence of other

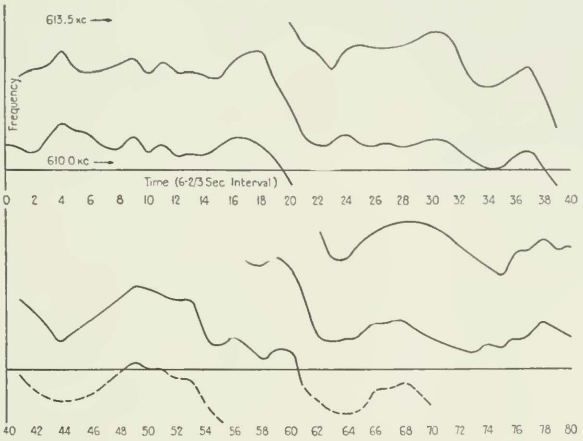


Fig. 20—Plotted curves which condense a long band fading record so as to bring out the frequency spacing interval of the selective fading

depressions in regular spacing beyond the scope of our pictures, for when one minimum slides out of sight another appears to take its place from the opposite side of the band. The minima traces shown in broken line were outside the record but were located by extrapolating the sections.

Other depressions of small amplitude appear to be superimposed upon the major changes but the present data appear inadequate to give reliable information concerning them. These minor depressions seem most evident during periods of rapid change.

The presence of these major minima in regular array bears a marked similarity to the familiar wave interference case in light and fits in very nicely with the theory detailed in previous paragraphs. Assume for a moment the simple case of two transmission paths producing such an effect and account for the difference in their lengths by presuming that one path follows more or less closely along the surface of the earth while the other seeks higher altitudes and in some fashion gets back to earth at the receiving station.

The mean frequency difference or spacing interval between successive minima for the records given in Fig. 20 is approximately 2,200 cycles. Therefore, the mean wave length difference in length of path from equation (5) is 277 wave lengths, or 136.5 kilometers.

It is evident that the errant waves following the second path must have been led a devious route. While this is about all the information which can be deduced directly from these data it is interesting to speculate further with the information along the lines of some of the theories which have been proposed to account for such wave deflections. For instance there is the Heaviside layer theory in which there is supposed to be a more or less well defined reflecting layer in the upper atmosphere. For this we would visualize our high altitude waves as proceeding in a straight line up to the layer, being reflected, and striking back to earth at the receiving station.

Since the distance from transmitter to receiver was 110 kilometers the length of the secondary path was 110+136.5 or 246.5 kilometers. By triangulation the height of the assumed reflecting layer may be determined as very nearly 110 kilometers or equal to the distance from transmitter to receiver, and the angle of incidence is 26.5 degrees.

As yet no positive information has been acquired concerning the variation of difference in length of two major night-time transmission paths with direct distance from the transmitter. If the path difference is due to reflection from an overhead layer, the expected relation by triangulation becomes quite simple.

$$\Delta d = \sqrt{\frac{y^2}{4} + h^2} - y.$$

When Δd is the difference in length of path, y is the direct distance and h is the vertical height of the layer.

An investigation of this relation would probably do much to prove or disprove the reflection theory.

At this point it is well to recall the results of earlier tests in which it was observed that single frequency waves separated by 1,000 cycles faded in approximately an inverse relation also indicating a spacing interval of about 2,000 cycles. The agreement of these earlier records is particularly noteworthy since about three weeks elapsed before the more detailed band fading records were made.

Fig. 20 shows a time variation in the frequency position of the minima which is explained as due to a variation in the difference of path length. If we indulge in further speculation along the line of layer phenomena we conclude that the reflecting layer is rising and falling. It is improbable that the whole layer would rise and fall

together so we conclude that undulations occur along its surface. These undulations in themselves would cause the length of path of the wave reflected toward the receiver to undergo a continual change. They would also introduce minor reflections from surfaces more distant than that responsible for the major effect which may be responsible for the more rapid, low amplitude fading which is usually superimposed upon the slow changes. Obviously, the character of the fading would in the event that it is caused by undulations along the reflecting layer, be determined by the amplitude and direction of movement over the surface.

If, on the other hand, we examine the possibilities of theories such as those proposed by Nichols and Schelleng, Larmor and others in which the action of free electrons in the atmosphere is invoked we might visualize the waves on the second path as following a curved trajectory. Or we might have the two sets of waves start off together, become split by double refraction and eventually come together again. Perhaps their planes of polarization will have been rotated. In fact it is possible to build up what appears, we must confess, to be a highly imaginary explanation in which the wave interference is accounted for not on the basis of any great difference in path length but by the assumption that the amount of rotation is such a function of frequency that a change of about 2,000 cycles adds or subtracts a complete rotation, and the further assumption that one set of waves has had its plane of polarization rotated through several more complete rotations than has the other. The synthetic possibilities are almost endless and we must wait upon further data more varied in character before the facts can be established. In the present investigation we have not attempted to determine the mechanism of the transmission medium except insofar as it could be inferred from the results of our tests which were aimed at finding out just how radio signals look after they have been subjected to a trip through this mechanism.

Returning to the solid band fading record illustrated in Fig. 18, let us form some conception of the appearance of this figure were it extended toward the much higher and lower frequencies using as a basis of this conception the supposition that the existing record is systematically distorted by wave interference. For a given rate of change in the physical difference in length of path, such as would be encountered in the simple reflection case, the rate of movement of the minima across the band fading pictures would vary directly with the frequency. Therefore, we can extend the narrow section shown in Fig. 18 to form a wide band fading record such as is shown in

Fig. 21, wherein we are looking down upon the distorted surface, the minima being traced by the light lines. Toward the short wave end of the band it is evident that a fading record for a single frequency represented, for example, by a section parallel to the time axis and perpendicular to the page, $a-a'$, would show rapid fading, while a similar record at the long wave end of the range as $b-b'$ would give slow amplitude changes. Such sections representing theoretical

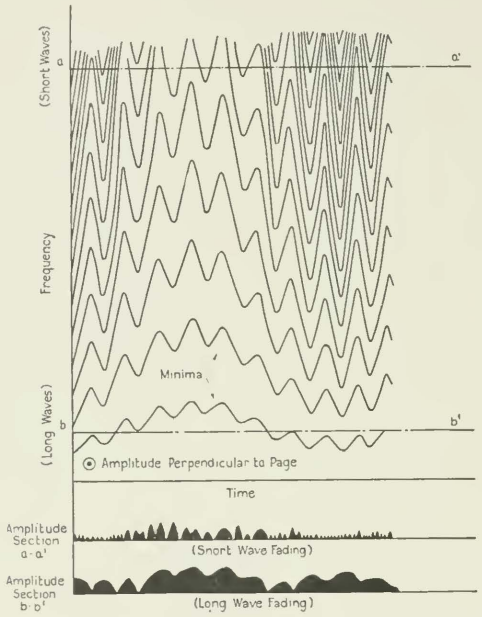


Fig. 21—Theoretical diagram obtained by extrapolating band fading records to show how the rapidity of fading might be expected to change with the wave-length

single frequency fading records are shown at the bottom of Fig. 21. The relative fading rates for long and short wave lengths as indicated by these idealized characteristics, are in accord with general experience

In describing the stepped-frequency method of obtaining band fading records allusion was made to distortion which might result from speeding up the process. Suppose that we were to use a very small rotating condenser in parallel with the main condenser of the transmitter oscillator for changing the frequency, and that this condenser were capable of changing the frequency sinusoidally about a mean value. Then we could represent the variation in frequency with time as is shown by the curve C_1 in (a) of Fig. 22. Now if the energy

transfer from transmitter to receiver takes place over two paths of different lengths one wave will constantly lag behind the other.

This lag may be measured as a time interval. In Fig. 23 are shown two waves, (a) and (b) of constant amplitude but with frequency

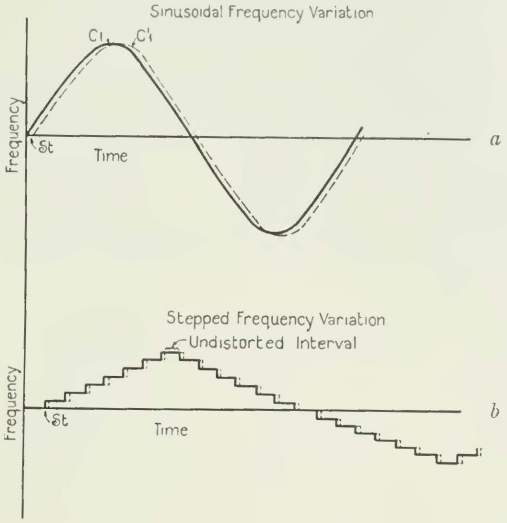


Fig. 22—Curves showing the relative effect of transmission time lag in sinusoidal and step-by-step methods of frequency variations

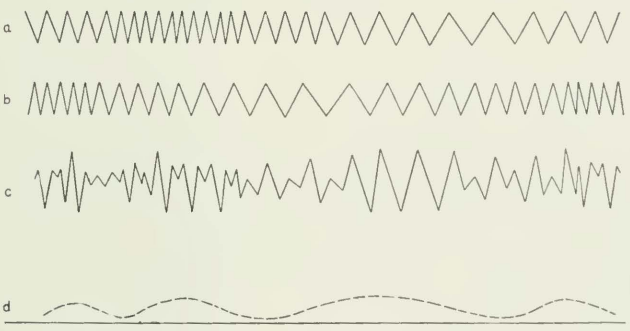


Fig. 23—Diagram showing the effect of frequency modulation

modulation. The wave (b) representing the indirect wave, it will be noticed, lags behind the direct wave represented by (a). The amount of this lag is determined by the difference in length of path and the transmission velocity. If we were to receive only one wave, as we should in the daytime, for example, we would find it to be a

constant amplitude field (providing the high-frequency characteristic of the receiver is flat over the range of frequency variation). But when two or more distinct paths exist, the combination at the receiver becomes complex. This is evident in curve (c) shown in Fig. 23 which is a direct summation of (a) and (b), and in (d) which is the envelope of (c). The amplitude is subjected to variations which did not exist at all in the original wave.

We might set up an *equivalent* effect right at the receiver by constructing two small local oscillators having the same characteristics as the transmitter oscillator. The two small rotating condensers would be driven by the same motor but the rotor of one would be shifted backward in phase relation to the other so as to simulate the case of transmission lag over the longer path. The relative frequency characteristics of the two may then be represented by curves C_1 and C_1' in (a) of Fig. 22.

The frequency of the signals arriving over devious paths at the receiver may be put in the form of an equation as,

$$F_1 = F_o + f \sin [r\ (t - d_1/V)], \tag{6a}$$

$$F_2 = F_o + f \sin [r\ (t - d_2/V)], \tag{6b}$$

wherein,

F_o = the mean frequency

f = one-half the total variation

$r = 2\pi$ times the frequency of rotation of the condenser

d = length of path

V = velocity of waves.

For a difference in length of path equal to 300 wave lengths at a frequency of 600,000 cycles per second, for example, the time lag of one wave behind the other will be equal to 300/600,000 second or 1/2000 second. The lag of one of the condensers behind the other in the "equivalent" case described above would be then for 30 cycles per second rotation of the condensers, 30/2000 times 360 degrees or 5.4 degrees. The lag of 5.4 degrees represents the lag of the condenser rotor so that the frequency lag will depend entirely upon the rate of change of frequency by the rotating condensers at any given instant.

Now to determine the resultant wave at the receiver we must know both amplitude and relative phase of the components arriving over the different paths. The amplitude will be constant, and we shall assume known, although it may actually follow slow changes with

attenuation or variations in length of path. The relative phase must be determined from equations (6a) and (6b). Knowing the frequency variation with time we may by integrating the following equation determine the phase relation at any time (t).

$$\theta_1 = \int_0^t 2\pi F_1 dt, \tag{7}$$

$$\theta_2 = \int_0^t 2\pi F_2 dt. \tag{8}$$

Substituting the general relation for F_1 and F_2 from equations (6a) and (6b) we have,

$$\theta_1 = \int_0^t F_o + f \sin r\,(t - d/V), \tag{9}$$

$$\theta_2 = \int_0^t F_o + f \sin r\,(t - d'/V). \tag{10}$$

Evidently the relative phase ($\Delta\theta$) will be the difference between these two giving,

$$\Delta\theta = \theta_1 - \theta_2 = 2\pi \int_0^t F_o\, dt + 2\pi \int_0^t f \sin r\,(t - d/V)\, dt \tag{11}$$

$$- 2\pi \int_0^t F_o\, dt - 2\pi \int_0^t f \sin r\,(t - d'/V) \tag{12}$$

which integrated reduces to the form,

$$\Delta\theta = \frac{2\pi f}{p}\,(\cos r\,t - 1)\,(\cos r\,d'/V - \cos r\,d/V + \sin rt\,(\sin rd'/V - \sin rd/V). \tag{13}$$

The equation is not in itself very illuminating, but what it tells us generally is that if we represent two frequency modulated waves travelling over paths of different lengths to a distant receiver by rotating vectors, these vectors are constantly shifting their relative position. The magnitude of the shift at any instant is given by the varying angle $\Delta\theta$. Due to a change in the angle included by the two vectors their resultant will undergo an amplitude change, the seriousness of which we will consider later.

Thus far in the discussion of frequency modulation by means of a rotating condenser we have assumed sinusoidal changes in frequency. The ordinary condenser departs considerably from such a performance. By considering the application of the integral equation for $\Delta\theta$ to such a case it will be recognized that the relative space posi-

tions of the vectors representing the direct and indirect waves will be subjected to changes at every point where the slope of the frequency-time curve departs from a simple sine relation. The degree of distortion due to the presence of such irregularities may be considerable.

In Fig. 24 are shown some samples of "wobbled" carrier frequency records obtained at Stamford, Connecticut. For these records the

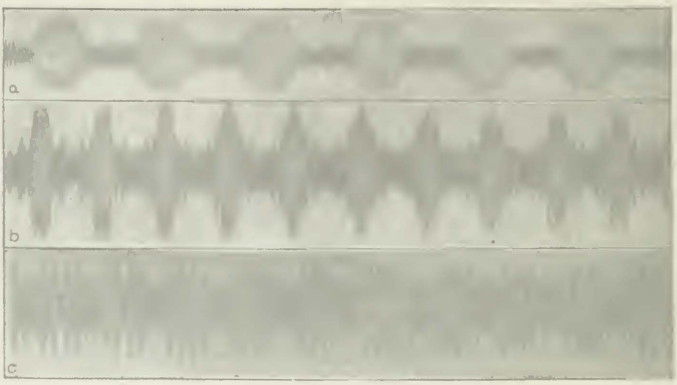


Fig. 24—Sample fast records showing distortion produced by intentional frequency modulation. *a* day record, *b* and *c* night records

carrier was wobbled at the rate of about 10 cycles per second. There is some uncertainty as to the range of frequency variation for these records although it was probably in the order of a few thousand cycles. By means of a constant frequency local oscillator the radio-frequency wave was stepped down in frequency to audio values which could be amplified and recorded.

The record (a) of Fig. 24 represents stable day-time reception. The record shows amplitude modulation due to the receiver characteristic alone. If the receiver were, as is desirable, capable of amplifying all the frequencies present in the received wave in the same ratio this record would be of constant width. In the subsequent examination of night records we must keep in mind the fact that the terminal apparatus is responsible for a certain part of the amplitude modulation. Its influence is readily recognizable.

The night-time records shown in (b) and (c) reveal a distinct distortion of the envelope aside from that present in the daytime record. Peaks appear and disappear within time intervals sometimes as short as a fraction of a second.

The record in Fig. 25 represents a slow picture of the changes shown in (b) and (c) of Fig. 24. If these wobbled frequency waves are studied carefully it will be noted that where a single peak stands at one moment there gradually comes in view another as if it were sliding from behind the first. The cycle length being about 1/10-second we may get some idea from this series of the rate at which the changes



Fig. 25—Sample slow record showing distortion produced by intentional frequency modulation. Night record

take place. The presence of so many peaks in these records is attributed in part to the fact that the rotating condenser used gave a frequency change which was far from a simple sinusoidal relation.

Let us now return to the stepped-frequency method of obtaining the band fading pictures and ascertain why it has certain advantages. In (b) of Fig. 22 is shown the "equivalent" characteristic for the stepped condenser. During 1/2000 of a second (for the conditions so far assumed) in each step distortion may occur due to transient conditions, but during the remainder of the quarter second assigned to each step (for the records so far taken) a steady state is reached. Thus, theoretically, distortion occurs only during about 1/500 of the step interval. In (b) of Fig. 22 the lag is greatly exaggerated for purposes of illustration. This means simply that we have maintained constant frequency for a sufficient length of time to establish, before taking our picture, a fixed interference condition over the region including transmitter and receiver at least.

Daytime Field Strength Distribution

Thus far we have been dealing with the unstable phenomena of night-time transmission. Our interest has been directed almost

entirely toward variations with time. While the presence of wave interference has been detected, and the movement of this interference effect across the frequency band has been recorded, little effort has been made to form a picture of such interference in its space relation. A discussion of similar stable, daytime phenomena is therefore not out of place, and particularly so in view of an evident relation of the fickle nocturnal interference phenomena to the steady states which follow the appearance of daylight.

In a previously published map of field strength distribution in New York City,* it was indicated that the congestion of high buildings

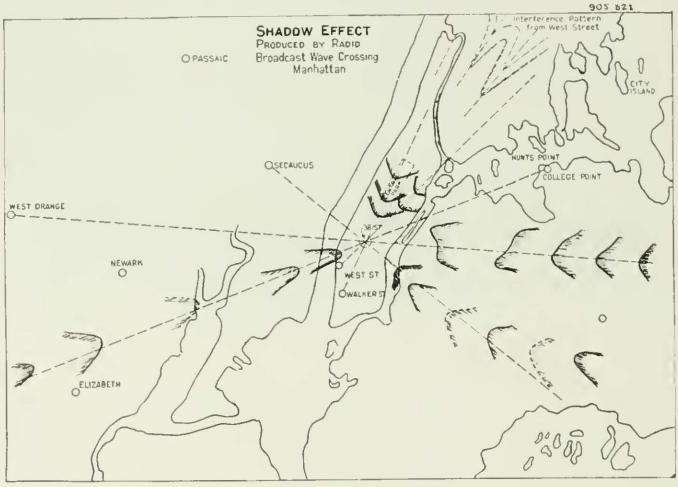


Fig. 26—Map showing location of radio obstruction on Manhattan Island as determined by the intersection of lines between various transmitting points and their corresponding shadows

just below Central Park cast a heavy shadow. More recently it has been determined from observations on a portable transmitter, set up at various points, that this building center is a consistent performer. The position of this obstruction is determined in Fig. 26 wherein only partial contours from maps for the indicated sites are given to prevent confusion. The intersection of these lines from transmitter to shadow, falls at approximately 38th Street in the vicinity of Sixth Avenue.

The dissipation of wave energy at such a point is probably the composite effect of many adjacent structures. Fig. 27 gives an elementary idea of how this can occur. The structures filling in

* See footnote 1.

each block are, of course, very well connected electrically by means of pipes, cables, etc., with those of adjacent blocks. Between each oscillating circuit (which is pictured as consisting of two buildings with earth connections) there exists a coupling which binds the whole system together more or less flexibly. Thus the obstacle offered by

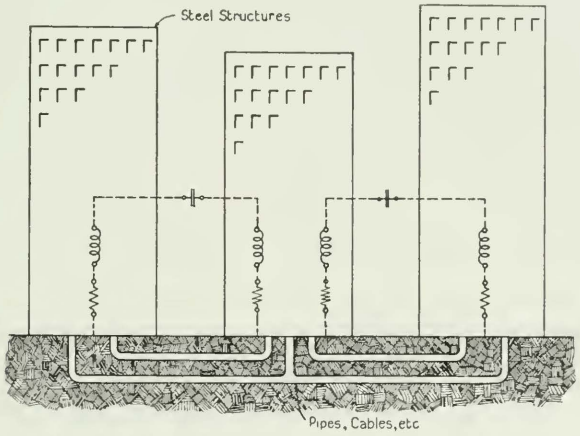


Fig. 27—Idealized picture of equivalent electrical circuit characteristics of high buildings

a group of buildings might be of a selective nature, and evidently its frequency characteristic may vary with direction.

Such an aggregate would, in addition to absorbing wave energy, produce a change in velocity or a refraction of the wave front. Some indication of such an effect will be discussed later. Before leaving the subject of shadows, however, let us get a physical picture of their significance.

From the transmitter a wave front expanding outward and upward encounters an obstruction which we shall assume is near the earth plane. The net result of this encounter is a weakening of the wave over an area near this plane, and probably a distortion of the energy-bearing fields. We might then imagine this shadow to be a tunnel-like region extending along the earth beyond the obstruction, and as having definite vertical as well as horizontal limits.

The aerial photograph of Manhattan and adjacent territory, shown on Fig. 28, will give a fairly clear idea of the conditions close to the transmitter. The major obstruction, the location of which has been previously described, is shown in its relation to the line of transmission toward the Riverhead and Stamford testing stations.



Fig. 28—Aerial photograph of Manhattan Island showing locations of transmitting station and obstructing high building area

Such barriers to wave travel, situated within a short distance from the source, seem, as we might expect, to have a more extensive and serious influence upon effective broadcast distribution than similar obstructions at greater distances.

It will be noticed that the obstruction falls very nearly upon the direct line from the transmitter to the Stamford testing station. This will also be evident later after an understanding of Fig. 29.

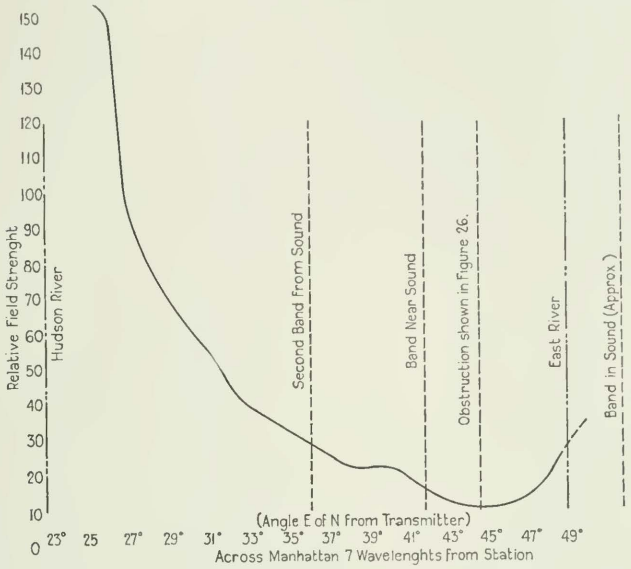


Fig. 29—Cross-section of radio shadow caused by high building area

wherein the position of the "Band Near Sound" represents also the bearing of the Stamford station. The Riverhead station is not directly in line with the major obstruction.

In certain sectors of the field strength contour map for station 2XB there appears to exist a kind of wavy displacement of the contour lines forming a partial pattern of peaks and depressions side by side. In general, this pattern must be differentiated from an ordinary shadow area. A remarkable example of this sort of field distribution is shown in Fig. 1 which is one section of a field strength survey made for station 2XB. These contours are based entirely upon daytime measurements, and represent a condition which is stable throughout the daylight period. Considerable difference in signal level is apparent within short distances across the direction of wave propagation. Two pronounced low signal channels extend ap-

proximately north-east across this region. These shift with change in frequency of the transmitted wave. Fig. 30 illustrates the space relations for such a movement. The full line curve shows a partial cross section of the contour map of Fig. 1 taken along a line approximately perpendicular to the direction of transmission 110 wave lengths from the transmitter. This represents relative field strength values for

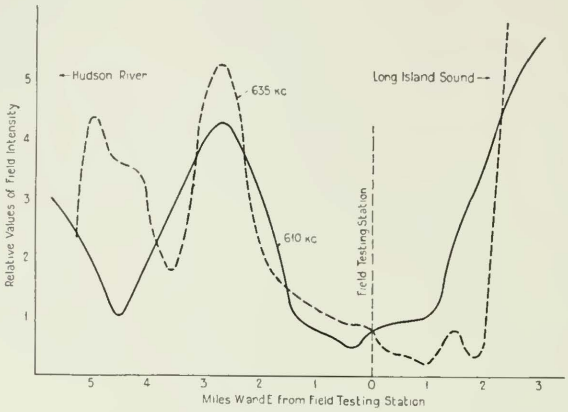


Fig 30—Cross-section of wave interference pattern showing change with frequency

610.0-kilocycle radiation. When the frequency is raised to 635.0 kilocycles, there occurs a movement of the peaks and depressions as is shown by the broken line of Fig. 30. Apparently the increased frequency causes these channels to be crowded together.

If we take sections of the field strength contour pattern in Fig. 1 and examine carefully the relative amplitude of peaks and depressions represented by these wavy lines we shall find that the ratio of field strength of the peaks to that in the depressions increases with distance from the transmitter. That is, the channels become more sharply defined as we move away from the transmitter. This ratio is shown approximately by the curves of Fig. 31. If these peaks or depressions were simple shadows they would maintain their relative values at a distance from the source or even tend to "heal" causing the ratio to fall rather than rise as is actually the case.

Within 14.4 wave-lengths (7.1 km.) of the transmitter the pattern, so apparent beyond 30 wave-lengths, merges into one deep shadow a cross-section of which is shown in Fig. 29. The abscissa of this curve is in degrees measured from the transmitter so that the center of the two most distinct low field strength channels extending north-

east may be inserted with their true radial relation. The two most evident in Fig. 1 are shown to be west of the line extending from transmitter through the center of the obstruction located in Fig. 26. The presence of Long Island Sound east of the geometrical center of the shadow has made an extensive survey of this section imprac-

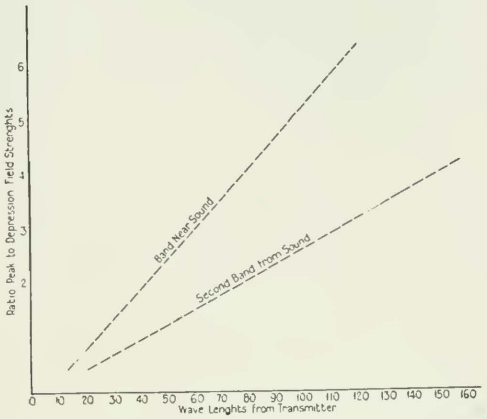


Fig. 31—Plot showing intensity of definition of wave interference pattern

tical. However, a single section taken across the Sound at about 90 wave-lengths from the station shows quite unquestionably the presence of a low channel about as indicated to the right of the obstruction designated in Fig. 29.

We have, therefore, a deep shadow with a more or less orderly array of maxima and minima within its limits. These maxima and minima grow more distinct at a distance from the transmitter, contrary to what we might expect for ordinary shadows. Furthermore, we find that they move as the frequency is changed. These facts lead to the belief that the phenomena in question are due to wave interference such as has already been described in connection with night-time fading, but characterized by very much smaller path differences. This daytime interference condition is fixed while we have seen that the nocturnal patterns appear to wander continually. To explain this more in detail let us return to the shadow and consider the phenomena which might accompany it in a little more detail.

The study of light has made available much information concerning the subject of wave interference. It is known, for instance, that the edges of shadows are not sharply discontinuous changes from light to darkness, but that a series of dark and light bands, called

diffraction fringes, are interposed between the full light and full dark areas. In our radio case the distance from the source to the obstruction and the dimensions of the obstruction are both very much smaller, in comparison with the wave length of the radiation, than for any ordinary case in light, but apparently the phenomenon is of the same general nature. By applying the ingenious principle of secondary sources used by Huyghens we might theoretically determine the distribution of the field beyond an obstruction placed in the path of the advancing radio waves. The basis of this principle is

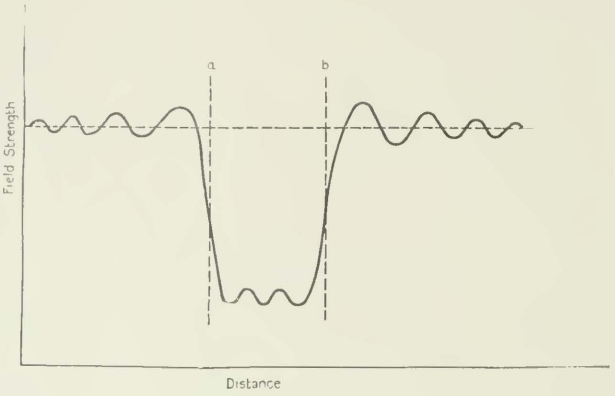


Fig. 32—Theoretical cross-section of radio shadow and associated wave interference pattern

the assumption that each elementary part of the advancing wave may be considered as a tiny transmitter. The effect at any point behind an obstruction, therefore, becomes the resultant effect, considering phase as well as amplitude, of the waves from all these miniature sources.

In Fig. 32 the region between vertical lines (a) and (b) represents the geometrical limits of the cross-section of a well defined shadow taken some distance behind the obstruction. An analysis of the resultant field using Huyghens' construction would show variations in intensity somewhat as represented by the full line. In other words the shadow will not be distinct but will have alternate maxima and minima within its geometrical limits and similar variations beyond the edges.

It is very likely, of course, that even in case the foregoing speculative analysis of the contour pattern extending north-east of 2XB is fundamentally correct, a great many other influences than that

of obstruction enter into the final field distribution. Relative attenuation of water and land appear to influence the distribution considerably though not as definitely as do steel structures close to the transmitter. Distinct minima appear both on the Hudson and on the Sound along radial lines extending from the transmitter.

Probably refraction of the wave front in passing across shore lines also enters into the shaping of this pattern.

Perhaps as good an elementary picture as any of the phenomena causing these patterns is that of a "dent" produced in the wave front by an encounter with a portion of New York City's impressive skyline. Since radio waves travel in a direction perpendicular to the plane containing the electric and magnetic fields, opposite sides of this "dent" would cross over one another with the result that an interference pattern would appear beyond the obstruction. An analogous situation exists when a water ripple passes a cluster of marsh grass which, damping its motion and retarding its progress causes part of the advancing front to converge and cross beyond the obstruction.

There is evidently a relation between day patterns such as have been discussed and night-time conditions. Just what this relation is offers some further opportunity for conjecture. In the first place quality distortion in transmission at night was, as previously explained, observed over parts of the region covered by the pattern shown in Fig. 1. The worst distortion seemed to be somewhat associated with the low field strength regions in this daylight survey. The distortion seemed also to be worse along the low channel extending in the direction of New Canaan, Conn., and beyond the 100-wave-length circle. It was particularly bad at a distance of some 140 wave-lengths from the station along this low channel where the field strength became so low in the daytime as to be unmeasurable with the set employed for the work. Accompanying the poor quality were fading and marked directional shifts.

Quality distortion though not so consistently severe at the Riverhead station as in the vicinity of Stamford was at times easily detectable by audible tests. Due to rapid attenuation of the radio waves traveling from the site of 2XB across Manhattan and the length of Long Island the field strength around Riverhead is generally low with higher levels north and south on the open waters of the Sound and Ocean respectively. Night-time fading at this point was representative of the variety which is usually found at distances of approximately one hundred miles from a broadcast transmitter.

The situation at Riverhead appears to be somewhat the same as

that which may exist over a large part of the broadcast area at a distance from the transmitter, while in the Westchester region we have an extreme and rather special circumstance. Field strength surveys have shown that there are indications of a daytime interference pattern over the Riverhead area but this pattern, such as it is, appears to be irregular and to lack the definition which makes the Westchester pattern so remarkable.

On the basis of the Westchester data alone we might build up a theory to the effect that night-time shifts of the stable daylight pattern were in some way responsible for quality distortion following the departure of daylight. Such a thought applied to the Riverhead case does not seem so reasonable since here the pattern is about one-quarter as distinct in terms of the ratio of maxima to minima values as the Westchester pattern. If, however, we presume that quality distortion may be expected in areas where daytime signals arrive *considerably attenuated* or so interfering as to simulate such an attenuated condition both situations are satisfied. After a consideration of the evidence at present available, such a conclusion seems attractive; that is, a daytime wave interference pattern alone is only an agency in night-time quality distortion in so far as its minima in combination with the general shadow effect are responsible for a low signal *directly* transmitted. Perhaps, in other words, the daytime field strength is a measure of *direct* night-time transmission, there existing in combination with this direct path at night a second, variable route of greater effective length. Probably close to the transmitter the "direct wave" is large compared to the "indirect" but shadows or interference may materially modify the ratio.

Night Distribution of Field Strength

By receiving simultaneously at several points the signal coming from a distant transmitter, it ought to be possible to detect the movement in space of these interference bands we have been discussing. The question immediately arises as to how far apart these distributed receivers can be placed without giving us an entirely discontinuous and misleading picture. For the first step toward recording space variations, in the vicinity of the Riverhead testing station, the receivers were spaced 1/16 wave length (30.5 meters), as illustrated in Fig. 33. It is necessary in making such determinations to transmit a single radio frequency, since we have already found that the interference bands for one component of a modulated wave are likely to be in a different position than those for another.

In order to receive and record the radio frequency wave it is, as has already been shown, convenient to use a local oscillator to beat it down to audible values. Since several oscillators for the separate sets are likely to produce mutual interference a common one was

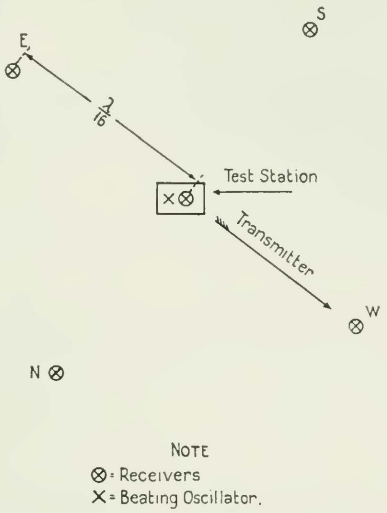


Fig. 33—Diagram showing space relation of receiving sets for special test



Fig. 34—Sample single-frequency fading record from spaced receiving sets

employed. This beating oscillator was situated at the testing station and the receiving antenna at this point was used as a radiator. In order to prevent overloading, the local receiver, the coupling to the receiver input coil was balanced to give a minimum of the local signal.

Fig. 34 is a sample of the record obtained. The continuous shadow band at the top represents the local receiver output. One oscillator

element was used for the other four receivers, their signals being recorded successively by a commutating device. Incidentally the interaction between these receivers was checked by observing the output of any one, while changes were made in the tuning of the others. The antenna was, however, so nearly aperiodic that no recognizable distortion or reradiation phenomena could be detected.

Fig. 35 illustrates compactly variations recorded by the oscillograph records (of which Fig. 34 is a sample), for a representative period of about five minutes. Even within the dimensions of 1/16

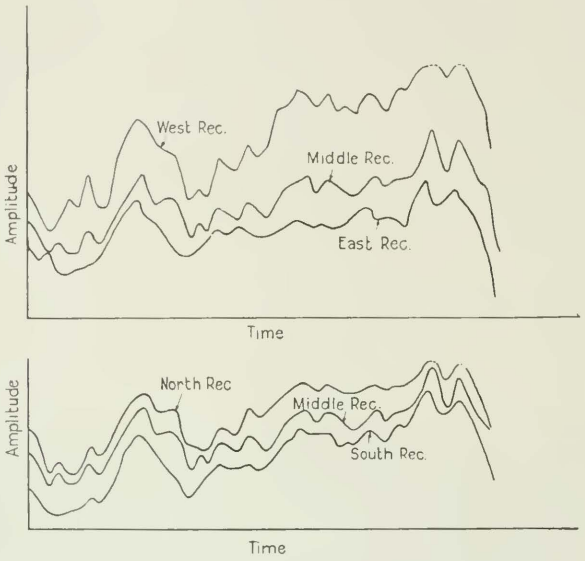


Fig. 35—Curves showing single-frequency fading on spaced receivers, condensed from long record

wave length there appears to exist transient field strength gradients in the direction of transmission. This is shown by a change in relative values, in the upper set of curves which represents field strength at points 1/16-wave length apart in the direction of transmission. The deviation is particularly noticeable in the relation between values for the local receiver and the "West receiver" which is in the direction of the transmitting station.

The lower set of curves, representing similar values across the line of transmission are much more nearly parallel. From the data so far obtained for the Riverhead testing site, it seems that transient nighttime field strength gradients are more generally evident in the direc-

tion of transmission than perpendicular to this direction. Upon these limited data one might be tempted to predict the presence of interference bands across the line of transmission.

The above discussion concerning space relation of field strengths has been included merely by way of contributing an additional bit of evidence to the theory that the erratic type of fading ordinarily

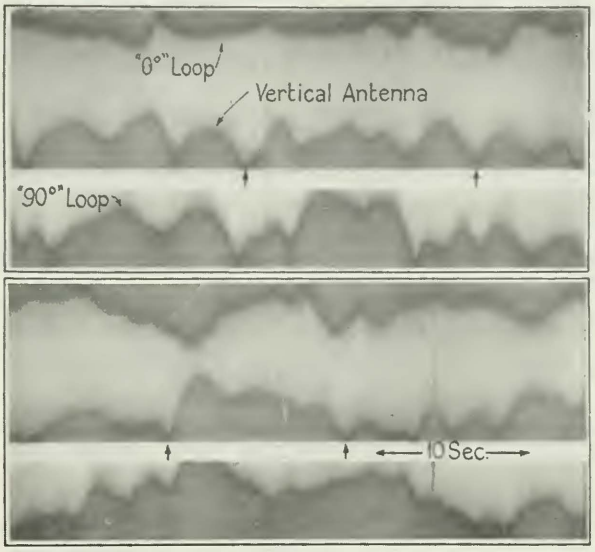


Fig. 36—Single-frequency fading record from vertical antenna and two-loop antenna crossed at right angles

experienced at night time is due to wave interference. The picture is very small in terms of wave lengths but considering its content, its very limits seem to imply wave interference rather than attenuation alone.

In connection with the wave interference theory thus far suggested as responsible for a major part of fading Fig. 36 is introduced as added evidence. The middle record of this group represents amplitude changes in the night-time reception of a carrier wave upon a vertical antenna. The upper and lower records represent the same for two loops turned at right angles to one another in the horizontal plane. By daytime tests the interaction of this combination was found to be negligible. Night-time fading recorded simultaneously for these three separate receivers occupying as nearly the same point in space as was possible, show that a high amplitude

signal may be coming in on both loops while the vertical antenna pick-up approaches zero. Several points of this kind are marked by arrows below the middle trace in Fig. 36.

There are at least two simple possibilities which might account for these relations. In case the wave approaches the receiving point from directly overhead, the vertical antenna would receive a "zero" signal while the loops would pick up an amount depending upon the state of polarization. If this be true, the records indicate a very rapid shift from the vertical direction of reception since the antenna minima are short lived most of them lasting at best a small fraction of a second.

On the basis of wave interference it is apparent that two waves approaching the receiving point in a 90-degree space phase relation and 180 degrees out-of-time phase could give a maximum signal on the two loops while that received on the vertical antenna was a minimum.

A compromise between these two viewpoints is probably a better guess than either one of them taken alone. That is, the existence of minima on the vertical antenna at the same moment that a strong signal is coming in on the loops is perhaps due to the interfering combination of waves having components in both the vertical and horizontal planes.

Quality Distortion

So far the data shown have been limited to the results of observations taken on special forms of transmission which are simplified for the purpose of clearly exposing the basic facts. We wish now to consider some of the more practical aspects of signal distortion. The first test which we made at our field test station was to record on slowly moving photographic paper tape and on the high speed film, the detected audio signal which resulted when the transmitter was modulated by a pure 264-cycle tone.

Fig. 37 is a sample of the general type of audio signal record obtained and Fig. 38 shows copies of the wave shape of the received signal, at particular times corresponding to the numbers of the oscillograms on the records in Fig. 37. The abrupt displacement of the timing trace indicates the point on the long record at which the snap-shot oscillogram was made. A peculiar characteristic of these records is the dark shadowy lines weaving back and forth through the band recording the complete signal. These dark lines correspond to the kinks in the wave shape shown in Fig. 38. As explained before, the darkening of the record is caused by the greater quantity of light

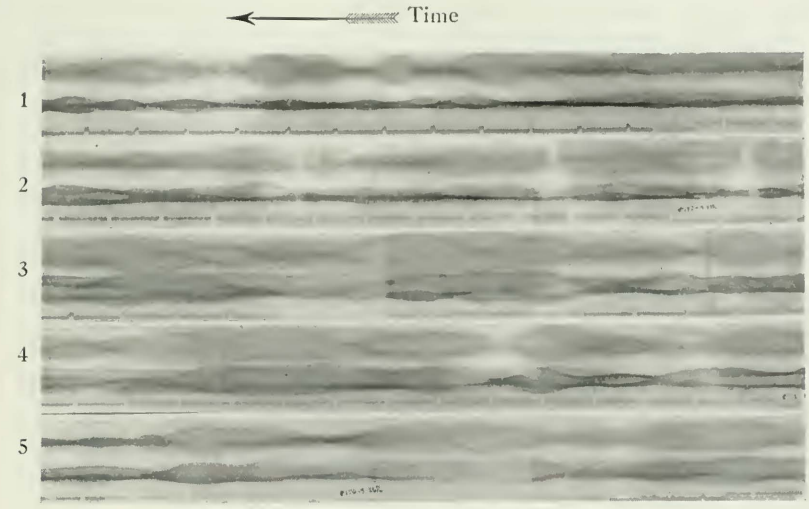


Fig. 37—Slow record of signal detected from tone modulated transmission showing the night-time distortion. Made at Stamford, Conn., May 15, 1924, 2:25 a.m. Upper trace signal from vertical antenna receiver and lower trace signal from loop antenna receiver, timing marks 2.6 seconds apart

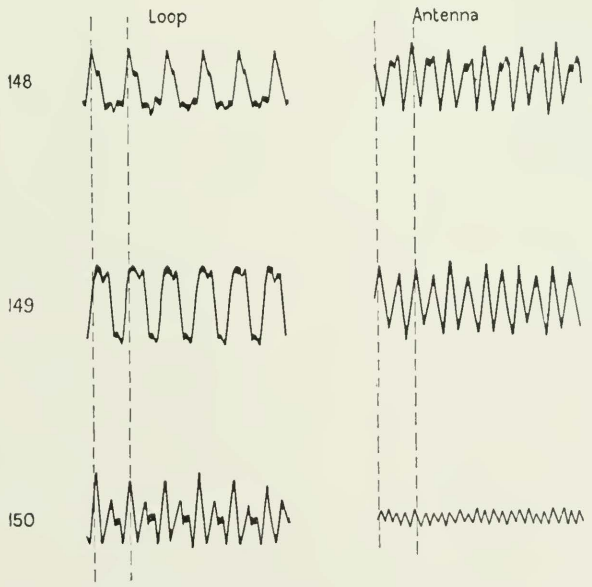


Fig. 38—Wave form of signals corresponding to numbered positions indicated on strips 2, 4, and 5, Fig. 37

affecting the record at these peak points. At the same time these observations were made, the wave shape of the signal rectified from the antenna current at the transmitter was recorded by an oscillograph. These oscillograms showed the signal to be free from distortion at the transmitter.

The weaving of these shadowy traces together with their width gives a record of the change in phase and amplitude of the irregularities in the wave shape of the signal. Although the wave shape of the signal is continually changing, it persists in substantially the same form for a great many cycles. Thus the record shows that, in the transmission of this simple tone modulated signal from the transmitting to the receiving antenna, it has been so modified that entirely new frequencies appear at the receiver. This receiver was shown by local tests to be free of any appreciable distortion within itself. While these new frequencies look like harmonics of the modulating tone in the snap-shot record it is obvious from the slow record that they are not true harmonics but that they differ from the harmonics by a very small amount and are incommensurable with the modulating tone since they undergo progressive but irregular phase changes with reference to it.

These records represent in a nutshell the signal distortion problem as it first presented itself to us. Our work then consisted in raveling out the complicated relations so that their nature could be ascertained and a theory of the causes established. In this paper, in the interest of clarity of presentation we have departed considerably from the actual order of the experimental work but at this point perhaps the actual order is best to follow for a moment.

With such a weird-looking distortion to analyze, and if possible eliminate, our first thought was as to whether the terminal apparatus might not involve unrecognized peculiarities which would be a contributing cause. Local tests and daytime tests of the receiving system absolved it from doubt and attention was focussed on the transmitting apparatus.

It was suspected that present day radio telephone transmitters leave something to be desired in regard to what we may call, for lack of a better term, their dynamic frequency stability. A very large percentage of the transmitters in use throughout the world today produce amplitude modulation of the carrier by the action of modulating tubes directly upon an oscillating tube circuit. It is to be expected that the cyclic changes in circuit conditions occurring at the modulating frequency will have some cyclic effect on the absolute frequency of the carrier and that this effect will be in the nature of a

wobbling or rapid shifting back and forth in frequency of the amplitude modulated carrier. In other words the carrier and side-bands, without change in their relative frequencies, would be subjected to "frequency modulation."

This sort of thing should be clearly distingushed from the slow wandering of frequency which, for instance, causes beat notes between carriers of different stations to drift gradually in pitch. What we have called "dynamic instability" is so rapid (being governed by the cyclic variations of the modulator) that it is difficult to observe by any aural method. Since the transmitter being used for our tests

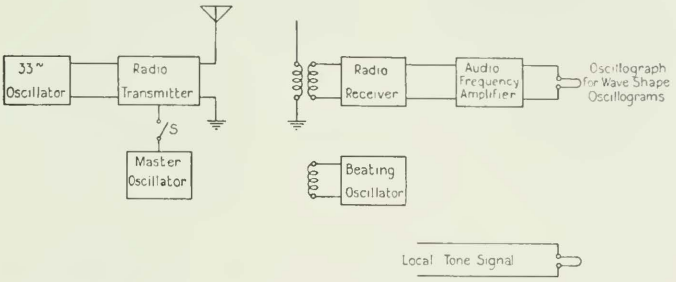


Fig. 39—Diagram of system used to measure frequency modulation

was a member of this almost universal class which employs modulating elements directly associated with the oscillator elements we determined to study this aspect of the transmission.

The following test was made to find out the extent of the frequency variation during the period of the modulating cycle. A schematic of the testing circuit arrangement is shown in Fig. 39. The plan was to modulate the carrier with 33 cycles, a tone so low in frequency that it would not be efficiently transmitted through the audio frequency amplifier connected to the output of the radio receiver. Then upon beating the received modulated carrier signal down to a frequency of about 1,000 cycles, an oscillogram of this signal would show a 1,000-cycle signal with a 33-cycle modulation in amplitude. Frequency modulation, if present, should then be easily discernible from the record. This experiment was made for daytime transmission and oscillograms (A) and (B) shown in Fig. 40 were obtained, one with the frequency of the beating oscillator greater than the carrier frequency, and the other with the beating oscillator frequency less than the carrier frequency. Both of these oscillograms show by the change in the frequency of the beat note signal

that frequency modulation occurs in the transmitter circuit. The frequency change is very apparent on the oscillograms when the lengths of one cycle at maximum and minimum amplitudes are compared. The reality of the effect is demonstrated in the two records, which by their difference show the reversal of the increased and decreased frequency points with reference to the modulation cycle when

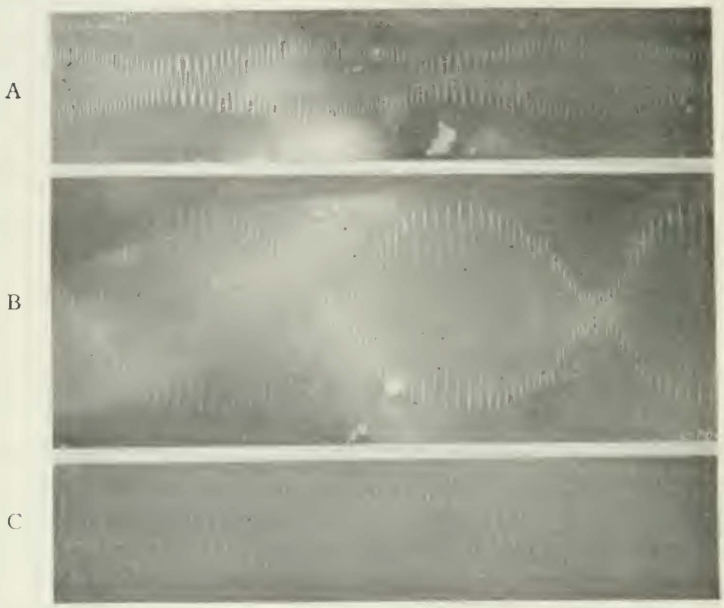


Fig. 40—Oscillograms showing frequency modulation accompanying amplitude modulation

the beating frequency is moved in frequency from one side of the carrier to the other.

The next step was to determine to what extent a stabilization of the carrier frequency to stop frequency modulation would affect the distortion of signals. True, master oscillator transmitters capable of giving the desired stability are not a new thing in the art. Several such transmitters were built by the Western Electric Company some years ago and used successfully in ship-to-shore radio telephone experiments [2] in which frequency stability was of considerable importance. To modify the ordinary broadcasting transmitter to in-

[2] See Fig. 1 and accompanying discussion in: Radio Extension of the Telephone System to Ships at Sea by H. W. Nichols and Lloyd Espenschied Proc. I. R. E., Vol. II No. 3.

clude this feature involves major mechanical changes and in order to provide a suitable arrangement for these tests the Bell Telephone Laboratories engineers merely added to the existing transmitter at station 2XB a temporary separate oscillator and high-frequency amplifier which could be connected to drive the oscillator tubes of the set as amplifiers. That this was free from frequency modulation is seen by comparing (C) of Fig. 40 with (A) and (B).

The transmission tests carried out with this arrangement yielded highly satisfactory results as indicated by a comparison of Fig. 41

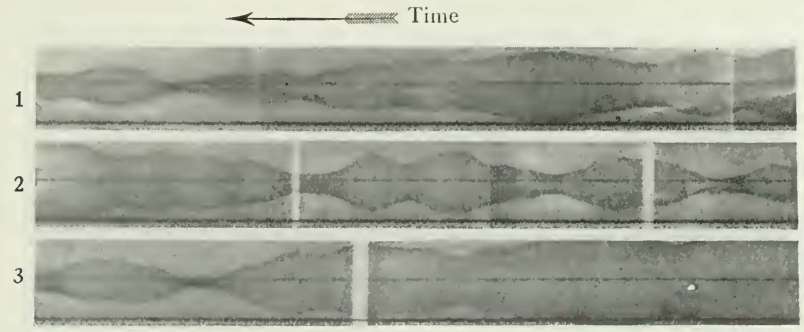


Fig. 41—Slow record of signal detected from tone modulated transmission with stabilized carrier showing reduction in distortion. Made at Stamford, Conn., Oct. 10, 1924, 3 a.m.

with Fig. 37. Fig. 41 like Fig. 37 is the detected result of a signal which started from the transmitter as a pure tone modulated signal, but it shows that much of the wave form distortion has disappeared, there remaining only a residuum which characteristically appears at the lower amplitudes of the signal. The probable cause of this residual effect will be discussed later. Tests of speech and music were concurrent with these findings. Using the normal transmitter, night-time transmission as received at the test stations was seriously distorted. When the stabilizing arrangement was employed this distortion was apparently eliminated except at the minima of fading.

Having arrived then at this practical result we wished to make further confirming tests, and tests to determine the whys and wherefores of the result. We have already detailed the more basic of these tests in previous sections of this paper and are now ready to consider the practical distortion records more carefully and to build up a theory to explain them.

The records shown in Fig. 42 are similar to the records in Fig. 37. They are shown here to illustrate the difference in the characteristics

of the wave form distortion variation that occurs from day to day. All these records were made at Stamford, Conn.

Strips 1 and 2—May 15, 1924—4:30 a.m.
Strips 3 and 4—Jan. 23, 1925—5:30 a.m.
Strips 5 and 6—Jan. 24, 1925—6:15 a.m.
Strips 7 and 8—Jan. 24, 1925—8:00 a.m.

There is a marked difference in the records obtained on January 23 and 24, which were made at the time an effort was being made to determine the effect of the solar eclipse on radio transmission. The peculiarly

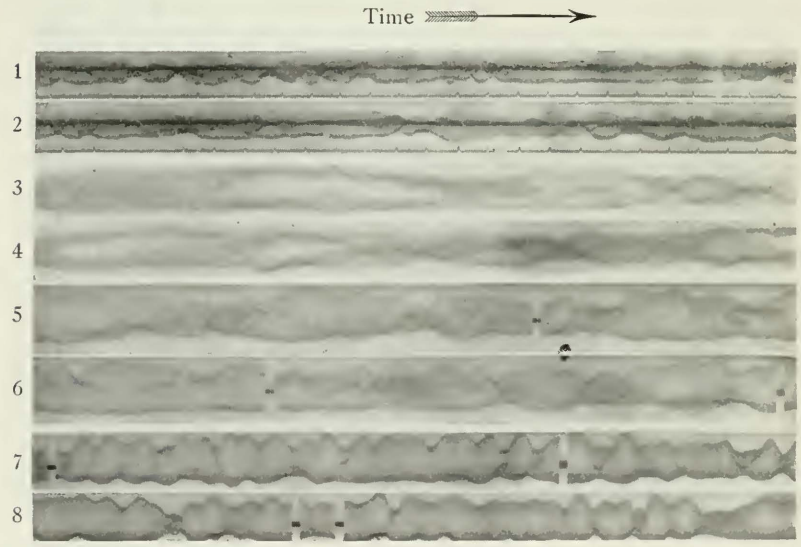


Fig. 42—Slow record of signal detected from tone modulated transmission taken on different days showing the changes in the character of the distortion

twisted appearance of the record obtained on January 24 is not very common in the records obtained. Most of the records have characteristics similar to those shown in Fig. 37. In the January 24 records there is a marked change in the characteristic configuration of the variation.

In order to obtain a record of the amount of wave form distortion resulting from frequency modulation present in the detected audio signal the circuit arrangement shown in Fig. 43 was used. This circuit was designed to analyze the wave form distortion when a 250-cycle

signal was used to modulate the carrier. Special precautions were taken to obtain a pure 250-cycle modulating tone. The wave shape of the signal detected from the carrier at the transmitter was frequently checked by observations with an oscillograph. The signals detected from the antenna current at the transmitter, both for the normal

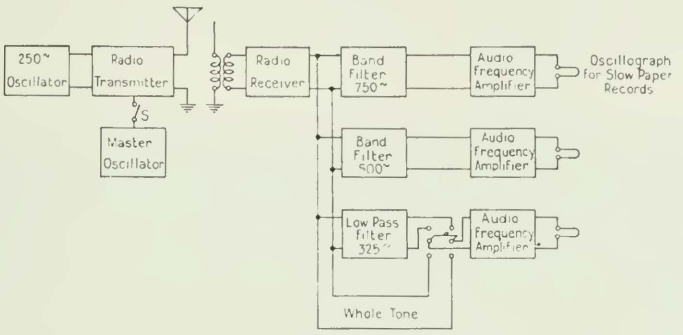


Fig. 43—Diagram of system used to obtain "Harmonic" analysis distortion records

transmitter with frequency modulation and for the stabilized carrier transmitter, were practically simple sine waves. The output circuit of the radio receiver was connected to a group of filters designed to transmit narrow bands of frequencies straddling the harmonics of 250-cycles.

While below we have referred to the frequencies passing these filters as "harmonics" it should be borne in mind that they are not necessarily *true* harmonics since they deviate very slightly from the true harmonic relation. The purpose of the test was to procure a record which would show at a glance the presence or absence of wave form distortion.

The input circuits of the filters were connected in parallel and the output circuits separately connected to the audio amplifiers arranged to operate the oscillograph elements. The input of one amplifier was arranged so that it could be switched either to the output of the filter passing 250-cycles or the output of the radio receiver. In this way a record could be obtained of either the whole tone from the receiver or only the 250-cycle component.

In Fig. 44, Strip 1 is a harmonic analysis record of the audio tone detected from the carrier and both side bands, transmitted with a stable carrier frequency. Strip 2 is a section of a record made a few minutes later when an unstabilized carrier was being used. On this record the lower trace is the 250-cycle component, the center trace the

500-cycle component, and the upper trace the 750-cycle component. The upper and lower traces have their zero lines at the edges of the strip. This record was made at Riverhead, L. I., April 30, 1925, at 3:33 a.m. Strip 2 is a section of a record made a few minutes later when an unstabilized carrier was being used.

The gain in the audio amplifiers connected to the outputs of the filters was adjusted to give nearly uniform transmission through the

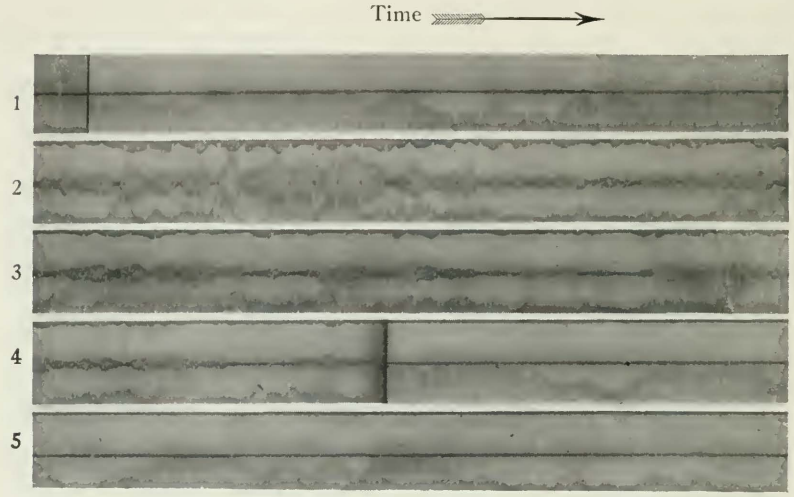


Fig. 44—Slow record made with system diagrammed in Fig. 43. Contrasting the distortion of detected tone transmitted by stabilized and unstabilized carrier frequency

receiving and recording apparatus for the frequencies recorded. Hence in these records the relative amplitudes of the fundamental and harmonics of the signal are directly comparable.

Strips 3, 4 and 5 in Fig. 44 are taken from a record made for the purpose of obtaining a comparison of the wave form distortion sustained by the detected audio signal transmitted by the normal transmitter with frequency modulation present and by a stable frequency transmitter. In each strip the lower trace is the whole tone from the output of the radio receiver, the middle trace the second harmonic (500 cycles) and the upper trace the third harmonic (750 cycles). Strip 3 and half of Strip 4 give the record obtained when the normal transmitter was used, and the remainder is the record obtained when the modified transmitter was used. There was a few minutes' difference in time between the ending of one transmitting condition to the beginning of the next during which the master oscillator control was switched

on at the transmitter. The receiving circuit was not changed during the making of this record, so that the results obtained from the two transmitters are directly comparable.

The record of the signal from the normal transmitter shows an abundance of second and third harmonics, at times equal in amplitude to that of the whole tone signal. The latter, of course, includes these harmonics. It will be noted also that dark line shadows run through the trace of the whole tone, indicating the presence of the wave form distortion. The signal from the stable frequency transmitter as shown by the record is practically free from wave form distortion. The trace of the whole tone is also free from any dark lines which would

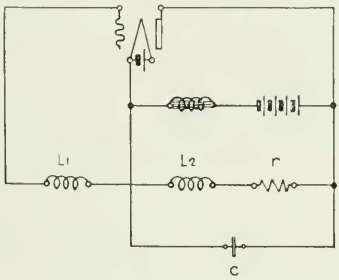


Fig. 45—Diagram of an oscillator circuit

indicate wave form distortion. This record is substantial evidence that a great deal of the wave form distortion may be eliminated when the carrier is stabilized. However, the selective fading still remains.

The selective fading we have already explained more or less satisfactorily and we find that it does not materially affect the wave form of audible frequencies transmitted by a modulated stabilized carrier unless its changes are more rapid than any we have recorded. The crippled state of originally perfect tone waves after they have been transmitted by an unstabilized carrier, we have just observed. Now let us consider the possible causes of this difference. The carrier stabilization referred to here, may we repeat, is not stabilization against slow variations in frequency from second to second or from hour to hour but rather against rapid variations within the cycle of the modulating frequency.

The reason for such changes over the modulating cycle is that the variation of the impedance of a vacuum tube across the oscillating circuit necessarily causes a variation in the nature period of the oscillation. As a simple case, the circuit in Fig. 45 is given.

H. J. Vander Bijl in his analysis[3] of this circuit gives the natural frequency of oscillation as

$$n = \frac{1}{2\pi}\sqrt{\frac{\left(1+\frac{r}{r_p}\right)}{L_2C}} \tag{14}$$

when r_p is the plate resistance and the remaining constants are given in the illustration.

Direct modulation by the usual method involves a cyclic change in the value of plate resistance. Hence, according to the above equation, there results a cyclic change in frequency which, though relatively small, becomes of the utmost importance when subjected to the peculiar phenomena of night-time transmission.

By making certain assumptions concerning the nature of frequency variation as amplitude modulation takes place, it is possible to work out distorted waves corresponding to various assumed wave interference conditions at the receiver. Perhaps the most simple and instructive means for producing these distorted waves is by a graphical method.

The equation for modulation of a high-frequency wave by a single tone may be written

$$e = (A + kA \cos vt) \sin pt \tag{15}$$

When A represents the unmodulated amplitude of the wave, k is a factor determined by degree of modulation, v is an angular velocity of the tone wave and p is the angular velocity of the high-frequency wave. The amplitude factor in this equation may be considered as a vector which is undergoing a change in length in accordance with the term included in the brackets. For the purpose of our analysis we shall include the angular velocity imparted to this vector by the last term in the above equation, since we are interested in the envelope of the resultant high-frequency wave at the receiver and the relative phase relations for two waves directly and indirectly transmitted combining to form this resultant. Since both carrier waves are of the same mean frequency only the relative position need be considered.

Now in our graphical determinations for the case of two transmission paths different in length, we represent the two effective fields by vectors varying in length in accordance with the amplitude factor of equation (15). However, due to the difference in length of path,

[3] "Thermionic Vacuum Tube," by Van der Bijl, page 274.

the changes in length of one vector will *lag the changes in length of the other* by an amount

$$\phi = v\ (\Delta t) \tag{16}$$

when Δt equals the difference in time of transmission over the two paths and v is the angular velocity of the modulating tone. This angle ϕ for 500-cycle modulation may according to the data thus far described, amount to more than 90 degrees at the receiving points selected for observation.

In addition to the lag in amplitude there will be a lag in frequency change over the frequency modulation cycle. This lag which has

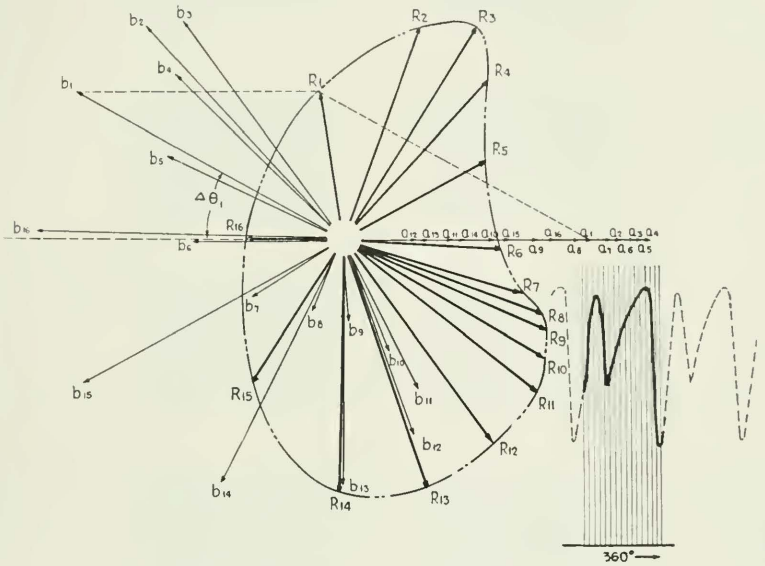


Fig. 46—Graphical method of synthesizing distorted wave forms caused by frequency modulation

already been shown in connection with the analysis of distortion in certain types of band fading records (see Fig. 22), becomes *a change in the relative phase angle* of the vectors under consideration. Thus our picture finally becomes one of two vectors changing in length, the changes in one continually lagging the changes in the other, the two vectors at the same time undergoing what we might term a relative angular wobble.

In Fig. 46 these relations are produced graphically. For our purposes we might assume that the vector representing one field is fixed and allow the other one to wobble the relative amount. At an

instant, for example, the directly transmitted field may be represented by a_1 in this figure. Assuming a difference in length of path, we may compute on the basis of the integral equation (13), the relative phase position of the vector representing the indirectly transmitted field b_1. The relative amplitude of this vector may also be determined by substituting $\Delta\phi$ in equation (15).

After establishing a sufficient number of vectors to represent the cyclic variation we may combine the respective components to obtain the resultant representative of the successive instants. These are

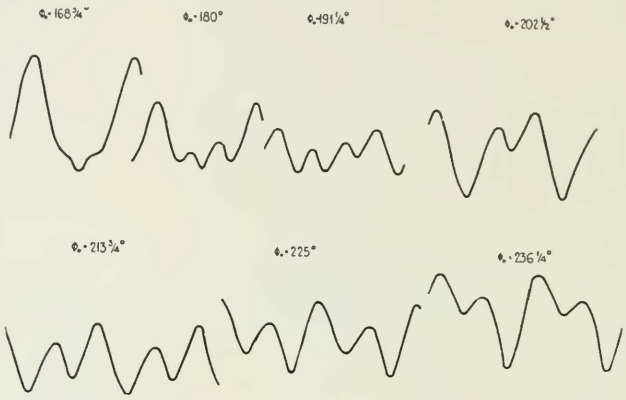


Fig. 47—Synthetic wave forms showing distortion due to frequency modulation

shown as R_1, R_2, R_3, etc., a broken line being drawn through their extremities to identify their positions. Now, if we plot these resultants as vertical ordinates in their successive time relation as shown on the lower right of Fig. 46, we have the envelope of the resultant wave at the receiver.

When the mean position of the two vectors (a) and (b) in Fig. 27 is 180 degrees separation, the signal is experiencing a fading minimum. When they are on the average in phase the amplitude is at a maximum. We can, therefore, trace a relation between quality distortion and fading by such an analysis, assuming a constant percentage modulation. Fig. 47 shows a series of high-frequency wave envelopes obtained by this method of graphic analysis. The mean vector relation is represented by ϕ_o, and for $\phi_o = 180$ degrees the fading may be considered at a minimum. The waves shown in Fig. 47 being envelopes of the high frequency will undergo certain changes in the process of detection. These, however, would only slightly modify the wave.

For purposes of comparison, a set of oscillograph pictures of representative received wave shapes is shown in Fig. 48. These represent the actual effect of night-time transmission with frequency modulation between 463 West Street, New York City and Stamford, Conn.; the modulating tone was a practically pure 264-cycle sinusoidal wave. The samples have been arranged in successive order

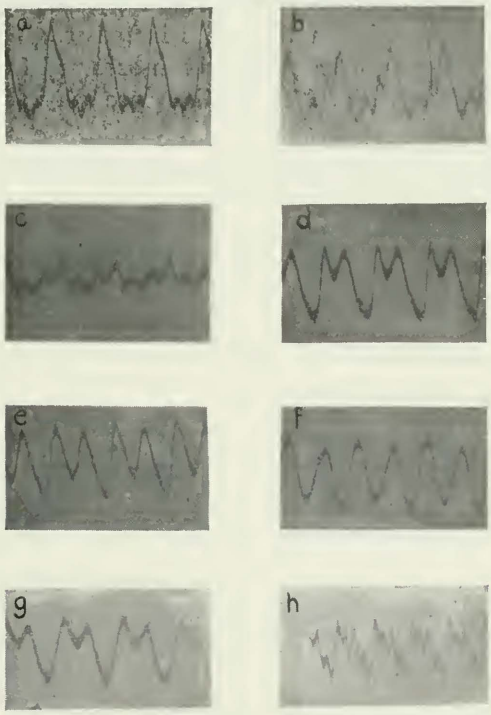


Fig. 48—Oscillograms showing actual wave forms with distortion resulting from frequency modulation

to correspond with the order shown in Fig. 47. There exists a striking similarity. Occasionally, however, the shapes predicted may depart considerably from those obtained experimentally. As an example of such a departure, the record (h) in Fig. 48 has been included. Such unusual samples may be due to a combination of waves arriving over more than two paths or it may be that the time variation of the frequency is far from the simple sinusoid which we have assumed. As a matter of fact, a critical mathematical treatment of this case shows that only an approximation of such a sinusoidal condition is

possible since as has been shown by Carson,[4] a frequency modulated wave of this character consists of an infinite series of fixed frequencies spaced at regular intervals either side of a "fundamental" carrier wave. Obviously only a small part of such a series could get out of the transmitter or into the receiver due to circuit selectivity. For the lower modulating frequencies, however, the approximation involved in the assumption of a simple sinusoidal variation is not far wrong since the amplitudes of these side frequency components fall off rapidly as their order in the series increases. While 150 wave lengths difference in path length has been assumed for the synthesis of the wave shapes in Fig. 47, this difference may according to the data obtained amount to much more than this.

It may well be asked why this frequency modulation, since it produces such marked distortion at night in certain places, does not also give rise to distortion by day or in locations where transmission is steady. A full answer to this question would be far from simple. But in brief it is because the carrier and side-bands shift in absolute frequency together as a unit so that their relative or difference frequencies which determine the audio signal remain unchanged. Another way to put it is that the detector operates on the envelope of the high-frequency signals and is blind to the frequencies contained within the envelope except insofar as they affect the latter. However, since frequency modulation appreciably widens the frequency band occupied by the radio signals it is to be expected that the tuned circuits in the receiver would have some reaction on those louder portions of the signal for which the amplitude modulation and therefore the frequency modulation is large. The perfection with which broadcast signals may be received under suitable conditions leads one to believe that this effect must be small.

Fading in Relation to Form of Transmission

It has been shown that serious wave form distortion of the reproduced signal may result if frequency modulation occurs with the amplitude modulation and the transmission is subjected to night-time conditions. This distortion from frequency modulation can be eliminated by stabilizing the carrier frequency. There remain some wave form distortion and the annoying amplitude changes caused by selective fading which is one of the most serious present day problems in radio transmission. Let us now consider the nature and cause of this

[4] See "Notes on the Theory of Modulation," by John R. Carson, Proc. Institute of Radio Engineers, February, 1922.

residual wave form distortion and some further consequences of selective fading under the assumption that there is no frequency modulation involved.

The process of detecting audio signals from radio frequency signals is, at least in its simpler aspects, well understood, but it may not be generally appreciated that the action is such that the detected signals may be greatly modified by changes in the relative amplitudes and phases of the carrier and side-band components such as may result from their transmission through the medium. That the amplitudes and phases of the carrier and side-band signals are not necessarily received in the same relation that existed as they left the transmitter has been pointed out earlier, in the discussion on selective fading.

The usual expression for a high-frequency carrier wave of frequency $p/2\pi$ modulated by a low-frequency wave of frequency $v/2\pi$ is

$$e = A[1 + a \sin (vt + \phi)] \sin pt$$

where A is the carrier amplitude, a, the percentage modulation and ϕ the starting phase of the modulating tone with reference to the carrier. Expanded into its components this becomes

$$e = \frac{A_1 a}{2} \cos (pt + vt + \phi_1) \qquad \text{(the upper side band)}$$

$$-\frac{A_2 a}{2} \cos (pt - vt - \phi_2) \qquad \text{(the lower side band)}$$

$$+A_3 \sin pt \qquad \text{(the carrier)}$$

where $\phi_1 = \phi_2 = \phi$ and $A_1 = A_2 = A_3 = A$ as the waves leave the transmitting antenna.

In the receiving set this function is squared by the action of the detector and, neglecting direct currents and frequencies above the audio range, the result is

$$\frac{a}{2} A_3 [A_1 \sin (vt + \phi_1) + A_2 \sin (vt + \phi_2)] + A_1 A_2 \frac{a^2}{4} \cos (2vt + \phi_1 + \phi_2) \quad (17)$$

of which the first term represents the fundamental frequency of the original modulating tone and the second term the second harmonic.

From this expression several conclusions can be immediately drawn. Due to the action of the detector there is always some slight wave form distortion as is evidenced by the presence in relatively small amplitude of the second harmonic. In the ordinary case this is negligible. The first term contains the carrier amplitude as a

factor but the second term does not. Thus, if selective fading erases the carrier at any time, reducing its amplitude to zero or a small value, the signal, represented by the fundamental tone, practically disappears, *even though the side-bands have not faded out*, and there remains only the harmonic. This is the residual distortion shown in Fig. 41 and which can often be heard during a fading out period. It is caused by the two side-bands beating together in the detector. We have here exposed a fundamental defect in the usual form of modulated signal transmission. The amplitude of the received signal is subject to all the whims of the carrier and to paraphrase freely an old saying we might remark that a signal is no stronger than its carrier. We may at once conclude that one way to reduce fading is to suppress the carrier and resupply a constant amplitude carrier at the receiving station.

Analyzing further the first term of the expression representing the detected signal, the first part of the bracketed portion results from beating together in the detector of the carrier and upper side-band and the second part from the carrier and lower side-band. It is clear that one of the side-bands may fade out completely and the other will still bring in the signal, provided the carrier is not also lost, with a phase shift to be sure but nevertheless not seriously reduced in amplitude. In telephony this kind of phase shift is relatively unimportant. Here we have an evident advantage in transmitting both side-bands since they support each other's frailties. But if the two side-bands suffer phase shifts in transmission, as we have earlier shown may be produced by wave interference, such that ϕ_1 and ϕ_2 differ by π radians or 180 degrees, the two components will cancel each other provided their amplitudes A_1 and A_2 remain equal. In other words all three components—carrier and both side-bands—may arrive at the receiver with full amplitude and yet no signal will be detected from them except a second harmonic component. This is obviously a disadvantage of transmitting both side-bands since, at such an instant, if one of them were eliminated the signal would reappear.

We conclude that there is, on the basis of such a brief analysis, not much to choose between single side-band and double side-band transmission when the carrier is transmitted also.

But if we wish to realize the advantages of carrier suppression a choice is not difficult. A carrier suppression system in which both side-bands are transmitted requires that the replacement of the carrier at the receiving station be done with almost absolute accuracy as to frequency and phase, a thing which involves very serious prac-

tical problems. On the other hand if but a single side-band is transmitted the difficulty is reduced to placing the carrier within a very few cycles of its correct position. The allowable departure will depend on a number of things but there is reason to believe that for high quality transmission it must be very small, perhaps no greater than two or three cycles.

With the single side-band carrier suppression method, invented by John R. Carson, the radiation is stripped down to the minimum which will fully transmit the telephonic signals and this reduces to a minimum the exposure of the signals to the ravages of selective fading. If the spacing interval of the fading is relatively narrow as in the cases we have examined hereinbefore, this form of transmission would not fade seriously in average volume but would be subjected to a continual changing of its frequency-amplitude characteristic, that is to say individual frequency components would fade progressively as the minima of the selective fading wandered back and forth across the frequency range encompassed by the single side-band. If the spacing interval of the fading were very large so that the minima were very broad of if some other, at present unexplored form of fading which covers a wide band at one time were acting, the signal would fade in average volume but the range of its variation would be only the square root of that of a carrier transmitted signal, since only the side-band would fade and the locally supplied carrier would remain unchanged.

The extent to which these theoretically drawn conclusions may be realized in practical application is yet to be determined but we have a few records bearing upon the matter which at least do not run contrary to them.

All of the transmission tests where the radio signal was beat with a local oscillator and the detected beat note observed, were equivalent to single frequency single side-band transmission with carrier suppression, the local oscillator functioning as the carrier suppressed at the transmitter. In this case, for which a number of records have already been shown, the detected signal is in proportion to the product of the amplitudes of beating oscillator and received radio signal. The phase of either does not affect the amplitude of the audio signal. Hence, the only important modification of the original signal is the variation in the amplitude resulting from selective fading.

Unfortunately we have no records in which a direct comparison is made between single side-band transmission with and without carrier suppression but the case can be visualized from the record shown in Fig. 12 or 13. Here each one of the frequencies recorded may be looked upon as a single side-band frequency which has been detected through

the agency of the resupplied carrier of the beating oscillator used to bring them down to audio-frequency. If now we were to take two of these frequencies shown on the record and multiply their amplitudes together at each point we would obtain the amplitude of the signal which would result if one of them were a single side-band and the other its accompanying carrier. It is obvious that the fading variations would thereby be increased in amplitude and rapidity.

In order to obtain a comprehensive picture of the relative advantages of radio transmission using a carrier and one side-band as compared

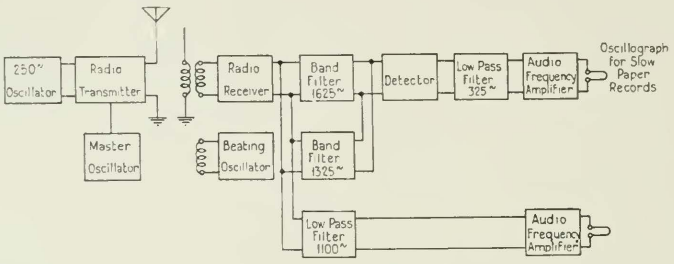


Fig. 49—Diagram of system used to obtain records of transmission with carrier and one side band and carrier and both side-bands

with the common practice of transmitting both side-bands, the following tests were made. The schematic diagram of the circuit arrangement is shown in Fig. 49. At the transmitter the carrier and both side-bands are transmitted and at the receiver they were selected out by means of filters in the manner previously explained. The signals from the filters corresponding to the carrier and lower side-band were applied to the input of a detector circuit and from its output the detected difference signal was selected by a low-pass filter. This signal was equivalent to that which would be received if only the carrier and one side-band were transmitted. From the output of the radio receiver a branch circuit goes to a low-pass filter which transmits only the signal detected from the carrier and both side bands, suppressing from this circuit the higher frequency signals corresponding to carrier and side-bands produced by the beating oscillator and received signals.

By making simultaneously a record of these two signals a direct comparison is obtained of the effect of selective fading on their amplitudes. Fig. 50 shows samples of several such records made at Riverhead, L. I. The modulating frequency for strips 1, 2 and 3 is 250-cycles, and for strips 4 and 5, 500-cycles. The record on strip 3 is shown on account of the peculiar characteristic of the signal fading, for considerable periods of time remaining at relatively low amplitude.

In these oscillograms the upper trace is the record of the signal from the carrier and both side-bands, and the lower trace the signal from the carrier and lower side-band.

These records illustrate by giving a graphic comparison the effect of the phase changes of the component signals in the case where the

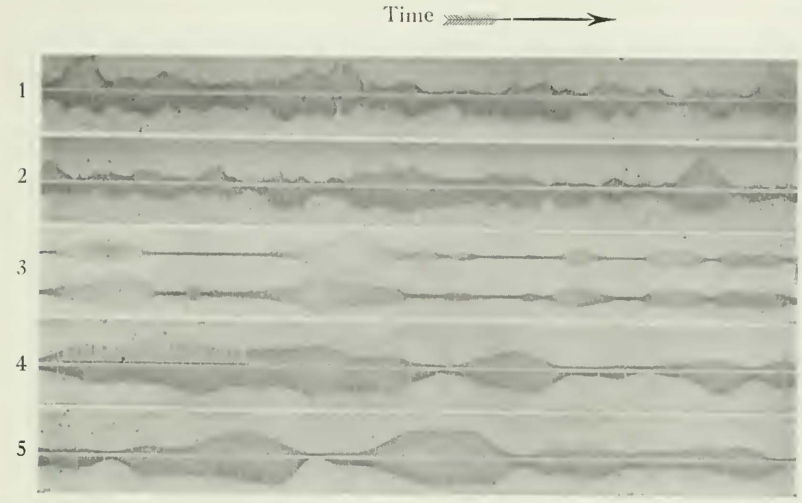


Fig. 50—Slow record comparing the signal detected from carrier and one side-band with signal detected from carrier and both side-bands. Made at Riverhead, L. I. Upper trace carrier + both side-bands, lower trace carrier + one side-band. Strips 1 and 2, July 22, 1925, 1:46 a.m. 250-cycle modulating tone. Strip 3, July 21, 1925, 3:10 a.m. 250-cycle modulating tone. Strips 4 and 5. July 23, 1925, 2:47 a.m., 500-cycle modulating tone

signal is detected from both side-bands. The amplitude of the signal from both side-bands in some instances is very small but appreciable amplitude is still indicated at the same instant for the signal from one side-band. This is explained as meaning that the side-band phases were such as to make the component signals 180 degrees out of phase after detection and that the amplitudes of the components were practically equal. The reverse situation is also observed where the amplitude of the signal detected from the lower side-band is zero and appreciable signal is recorded for the case where both side-bands are used. This is interpreted to mean that the side-band signal was eliminated by selective fading. In this event it was, of course, not contributing to the signal which was detected from both side-band signals. The recorded signal comes from the other side-band which evidently was not eliminated at that instant by selective fading.

Visual observations made with the cathode ray oscillograph, which unfortunately furnishes no permanent record of transient effects, confirmed the strip records in regard to the reality of there being side-band phase variations. From equation (17), it is seen that if these variations occur the fundamental of the detected tone signal at the receiver will not bear a fixed phase relation to that detected from the transmitting antenna current while if there are no such changes the phase between these two tones would remain constant. The locally detected tone and the tone detected from the transmitting antenna current and brought to the receiving station over telephone wires, were applied to the two pairs of deflecting plates in the cathode ray oscillograph. Since the deflections caused by these two pairs of plates are at right angles to each other the resulting Lissajous figure from two sine waves of the same frequency will be a slanting line, an ellipse or a circle depending on their phase and amplitude relation. The actual figures were observed to change progressively through this range of shapes, the changes following roughly the magnitude and rapidity of the fading. The effect of amplitude changes on such figures is quite distinct from the effect of phase changes and there was no difficulty in separating out the evidence of large phase changes.

Considering only the above theories and facts there appears to be a reasonable basis for a conclusion that the best form of radio transmission for use in broadcasting is single side-band with carrier suppression. But on practical grounds we do not believe such a conclusion is justified. The fading and distortions which we have made much of in the preceding pages are not experienced by the majority of broadcast listeners when they listen to local stations. To require these listeners to provide themselves with more complicated and expensive receivers, simply to allow more distant or less favorably situated listeners to obtain better reception, seems neither reasonable nor desirable. The art offers several other possible avenues toward improvement much less difficult of application and it must be remembered that radio broadcasting is already reaching a degree of standardization and a volume of existing receiving equipment which rules that changes must come slowly and without serious prejudice to the existing order.

Conclusions

Subject to the limitations imposed by the scope of our investigations the following conclusions may be drawn:

Fading can be quite sharply selective as to frequency and the evidence points toward wave interference as the cause.

The evidence for wave interference indicates that some of the energy of received signals reaches its destinations by a circuitous route and suggests that this route is by way of upper atmospheric regions.

Quality distortion may result from dynamic instability of the transmitter.

Fixed wave interference patterns in connection with shadows sometimes exist in daytime transmission.

Abstracts of Bell System Technical Papers Not Appearing in this Journal

New Methods and Apparatus for Testing the Acuity of Hearing. HARVEY FLETCHER.[1] This paper presented before the American Otological Society, classifies hearing tests in four groups according to their purpose.

1. Industrial or those made for determining the fitness of a candidate for employment. In certain types of work it is particularly important that a prospective employee meet a definite requirement for acuity of hearing. Tests made in the army and navy for various branches of service are conspicuous examples of this kind of test.

2. Educational or those made for determining the degree of hearing of school children both in the public schools and in the schools for the deaf for the special purpose of determining the proper methods to be used in their education.

3. Clinical or those made for assisting the physician to make a proper diagnosis of the cause of deafness.

4. Research or those made to determine new facts about both normal and abnormal hearing.

It is highly desirable that a single scale be used for representing the degree of hearing which is independent of the method used and which has a general application to the four purposes enumerated. Such a scale is proposed and it is shown how the commonly made voice test, watch tick, acoumeter, coin click and tuning fork tests can be expressed in terms of hearing loss units on this scale.

The paper is concluded by summarizing the different methods for testing the acuity of hearing which are as follows: (1) voice tests, (2) phonograph audiometer, (3) hearing loss for speech calculated from audiogram, which audiogram may be obtained in three ways, (a) tuning forks (constant initial amplitude), (b) tuning forks (comparison with hearing of tester), (c) pitch range audiometer.

The Relation Between the Loudness of a Sound and Its Physical Stimulus. J. C. STEINBERG.[2] Experiments with many types of sounds have shown that the loudness of a sound is a function of its

[1] *The Laryngoscope*, Vol. XXXV, No. 7, July, 1925.

[2] *Physical Review*, Vol. 26, pp. 507–523, Oct., 1925.

energy frequency spectrum and its level above the threshold of hearing and that if this relationship be represented as

$$L = \frac{10}{3} \log_{10} \left[\sum_{i=1}^{i=k} (W_i P_i)^{2r} \right]^{r^2},$$

sounds whose calculated values L are equal will appear equally loud to the average normal ear. P_i is the r.m.s. pressure of the ith component of the sound wave. The weight and root factors W and r, respectively, are functions of the sensation level, which is synonymous with the term loudness as formerly used and is defined as

$$S = 10 \log_{10} \left[\sum_{1}^{k} P_i^2 \Big/ \sum_{1}^{k} P_{oi}^2 \right]$$

where P_{oi} is the r.m.s. pressure of the ith component when the complex sound is at the threshold of hearing. In case the components in a narrow band of frequencies Δn are not resolved their energy must be integrated to obtain the energy of the equivalent single component. The root factor r is inversely proportional to the ratio of the minimum perceptible increase in energy to the total energy. For intensities near the threshold, the weight factors are equal to the reciprocals of the minimum audible pressures. Curves are given showing the values for W for various frequencies at various sensation levels, also the values of r as a function of S. As the intensity is increased the weight factors give greater weight to the lower frequencies; hence, even though the amplitude of the sound wave be increased without distortion, the ear will perceive both an increase and a distortion. This effect is due to the non-linearity of the ear.

Binaural Beats. C. E. LANE.[3] By introducing a tone of frequency f into one ear and another tone of frequency $f+N$ into the opposite ear, where N is less than 5 or 6 cycles, two kinds of binaural beats are obtained. Objective binaural beats are heard for most values of f within the audible frequency range, provided there is the proper difference in amplitude between the two tones. For telephone receivers as sound sources, this difference for best beats is about 55 TU and for the same receivers supplied with sponge-rubber cushions about 62 TU. These beats are heard because the louder tone is conducted through the head to the ear of the weaker tone and the two tones there are about equally loud. Subjective binaural beats are heard for frequencies below 800 or 1,000 cycles when the tones at the

[3] *Physical Review*, Vol. 26, No. 3, Sept., 1925.

two ears have about the same amplitudes, differing by not more than 25 TU. Data obtained with 22 observers are summarized. The evidence indicates that these beats are not due to cross conduction but are of central origin and the result of the sense of binaural localization of sound by phase. If the beats are slow (less than 1 per sec.) they are generally recognized as an alternate right and left localization, though some observers may report one or more intensity maxima during the beat cycle. Such maxima are explained as the result of one's interpreting the sound as louder when localization is more definite. Fast beats (more than 1 per sec.) are generally recognized as an intensity fluctuation. They are explained by assuming that the sound appears louder when the phase relations are such that it is normally best localized in the position toward which the attention is directed. This explanation is supported by observations made with a constant source rotating around the head of a listener.

Effect of Tension Upon Magnetization and Magnetic Hysteresis in Permalloy. O. E. Buckley and L. W. McKeehan.[4] Wires of five nickel-iron alloys containing 45, 65, 78.5, 81 and 84 per cent. Ni, 60 cm. long and 0.1 cm. in diameter, were studied by a ballistic method, for tensions up to 10,000 lb. per in.^{2} and fields up to saturation (10 to 20 gauss). Permalloy with 81 per cent. Ni is nearly indifferent to tension in its magnetic behavior; permalloy with less nickel is more easily magnetized and has less hysteresis when under tension, while 84 per cent. permalloy is more difficultly magnetized and has greater hysteresis when under tension. The saturation values are independent of the tension. In 78.5 per cent. permalloy, under a tension of 3,560 lb. per in.^{2}, saturation is reached at only 2 gauss (and is practically complete at 0.2 gauss) and the hysteresis loss is only 80 ergs per cm.^{3} per cycle, so small that it may be regarded as due to slight inhomogeneity rather than to any essential features of the magnetization process. *Relation to crystal orientation.* X-ray examination proves that this abnormally low loss is not due to any peculiar orientation of the crystal axes as the crystals are found to be oriented at random. Magnetostriction behavior can be deduced from these results. Above 81 per cent. Ni, permalloy contracts like Ni while below 81 per cent. Ni, permalloy expands like Fe.

Demagnetizing factor for a wire with a length 600 times the diameter, was determined experimentally and found to vary from a maximum of 1.6×10^{-4} to a low value, the changes being like these previously described by Benedicks for iron.

[4] *Physical Review*, Vol. 26, No. 2, Aug., 1925.

A Contribution to the Theory of Ferromagnetism. L. W. McKeehan.[5] *Relation of permeability and hysteresis to atomic magnetostriction.*—In permalloy, it has been found that magnetostriction changes sign at about 81 per cent. Ni, hysteresis losses can be made vanishingly small near this composition, and these effects are not due to the special alignment of crystals. It is suggested that in every ferromagnetic material the process of magnetization involves (1) intra-atomic changes, presumably changes in the orientation of electron orbits, governed by quantum dynamics and independent of environment; and (2) inter-atomic changes (stresses and strains). The interdependence of the inter-atomic changes and the intra-atomic changes is conveniently described as atomic magnetostriction. On this view, hysteresis loss and magnetic hardness are due to the energy required to produce, in succession, the local deformations associated with changes in the magnetization of single atoms or small groups of atoms. High initial permeability and low hysteresis loss in permalloy are explained as resulting from locally compensatory atomic magnetostrictions of the nickel and iron atoms in small groups. The fundamental differences in the magnetic behavior of Fe, Ni and Co are attributed to differences in their atomic magnetostrictions. Other differences are attributed to differences in the mechanical properties which alter the energy expended when atomic magnetostriction takes place.

Induction from Street Lighting Circuits: Effects on Telephone Circuits. R. G. McCurdy.[6] Synopsis.—This paper discusses series street lighting circuits from the point of view of their relations to nearby telephone circuits. These lighting circuits often have a much greater inductive influence in proportion to the amount of power transmitted than have most other types of power distribution or transmission circuits. This is due to the relatively large distortion in wave shape of voltage and current on certain types of these lighting circuits, and to the unbalanced voltages to ground which occur with series layouts. Three general types of lighting circuits are discussed. These are a-c, arc circuits, d-c, arc circuits supplied by mercury arc rectifiers, and alternating-current incandescent circuits. Of these, the incandescent type of circuit, in which the lamps are equipped with individual series transformers or auto-transformers, is the most important in this respect. Measures for reducing interference from these circuits are discussed.

[5] *Physical Review*, Vol. 26, No. 2, Aug., 1925.
[6] *A. I. E. E. Journal*, Vol. 44, pp. 1088–1094, Oct., 1925.

Power Distribution and Telephone Circuits. Inductive and Physical Relations. H. M. Trueblood and D. I. Cone.[7] Consideration of the relation between power distribution and telephone systems is naturally involved in the comprehensive review of the problems of the rapidly expanding power distribution networks in this country. Pending the completion of studies now being actively carried on in this comprehensive review, a preliminary and qualitative discussion is given.

Situations of exposure fall into three groups determined by the character of the area served. (1) "downtown" districts; (2) residential urban districts; (3) rural districts. The major problems arise in the second group. A wide variety of arrangements characterize both systems, and require consideration.

Among technical features, coefficients of induction for close exposures, shielding action of metallic cable sheaths for both power and telephone circuits, and "ground potential" effects, are distinctive problems. Where both classes of circuits are in cable with suitable precautions as to grounding, interference is rarely to be anticipated.

Noise induction from power-distribution circuits is chiefly from residuals, which occur on single-phase branches of polyphase circuits, or where triple harmonics or load-current unbalances are introduced by grounding neutrals, or where admittances to ground of phase wires are unequal. Residual currents are largest in systems having multiple-grounded neutrals, both load currents and triple harmonics occurring. Approximate resonance at triple harmonic frequencies between the inductance of station apparatus and power cable capacitance has characterized several situations. Various single, two and three-phase arrangements are compared from the induction standpoint.

The closely related matter of unbalances in the telephone plant is briefly discussed.

[7] *Journal of the A. I. E. E.*, Vol. XLIV, No. 12, Dec., 1925.

Contributors to this Issue

Bancroft Gherardi, M.E., M.M.E., Cornell University. Engineering assistant, 1895–09; traffic engineer, 1899, New York Telephone Company; chief engineer, New York and New Jersey Telephone Company, 1900–06; assistant chief engineer, New York Telephone Company, and New York and New Jersey Telephone Company, 1906–07; equipment engineer, American Telephone and Telegraph Company, 1907–09; engineer of plant, 1909–18; acting chief engineer, 1918–19; chief engineer, 1919–20; vice-president and chief engineer, 1920—. Mr. Gherardi's work in the field of telephony is too well known to require comment.

Robert W. King, A.B., Cornell University, 1912; Ph.D., 1915; assistant and instructor in physics, Cornell, 1913–17; Engineering Department of the Western Electric Company, 1917–20; Department of Development and Research, American Telephone and Telegraph Company, 1920–21; Information Department, 1921—.

Walter A. Shewhart, A.B., University of Illinois, 1913; A.M., 1914; Ph.D., University of California, 1917; Engineering Department, Western Electric Company, 1918–24; Bell Telephone Laboratories, Inc., 1925—. Mr. Shewhart has been engaged in the study of the relationship between the microphonic and physicochemical properties of carbon.

Harvey Fletcher, B.S., Brigham Young, 1907; Ph.D., Chicago, 1911; instructor of physics, Brigham Young, 1907–08; Chicago, 1909–10; Professor, Brigham Young, 1911–16; Engineering Department, Western Electric Company, 1916–24; Bell Telephone Laboratories, Inc., 1925—. During recent years, Dr. Fletcher has conducted extensive investigations in the fields of speech and audition.

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 many papers on theory of electric circuits and electric wave propagation.

KARL K. DARROW, S.B., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., in physics and mathematics, University of Chicago, 1917; Engineering Department, Western Electric Company, 1917–24; Bell Telephone Laboratories, Inc., 1925—. Mr. Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

RALPH BOWN, M.E., 1913, M.M.E., 1915, Ph.D., 1917, Cornell University, Captain Signal Corps, U. S. Army, 1917–19; Department of Development and Research, American Telephone and Telegraph Company, 1919—. Mr. Bown has been in charge of work relating to radio transmission development problems.

DELOSS K. MARTIN, B.S., Polytechnical College of Engineering, 1920; U. S. Navy, 1918–1919; Department of Development and Research, American Telephone and Telegraph Company, 1919—. Mr. Martin's work has related particularly to radio broadcast transmission.

RALPH K. POTTER, B.S., Whitman College, 1917; E.E., Columbia University, 1923; U. S. Army, 1917–19; Department of Development and Research, American Telephone and Telegraph Company, 1923—. Mr. Potter has been engaged in experimental work relating to radio transmission phenomena.

The Bell System Technical Journal

April, 1926

Development and Application of Loading for Telephone Circuits [1]

By THOMAS SHAW and WILLIAM FONDILLER

SYNOPSIS: A review of the art of loading telephone circuits as practised in the United States. The introductory section briefly reviews the theory of coil loading, and summarizes the principal characteristics of the first

CORRECTION SLIP FOR ISSUE OF JANUARY, 1926

Page 172: Equation should read

$$\Delta d = \sqrt{y^2+4h^2} - y.$$

Page 177: Equations (9), (10), (11) and (13) should read

$$\Theta_1 = 2\pi \int_0^t [F_o + f \sin r\,(t - d_1/V)]\,dt, \qquad (9)$$

$$\Theta_2 = 2\pi \int_0^t [F_o + f \sin r\,(t - d_2/V)]\,dt. \qquad (10)$$

$$\Delta\Theta = \Theta_1 - \Theta_2 = 2\pi \int_0^t F_o\,dt + 2\pi \int_0^t f \sin r\,(t - d_1/V)\,dt$$
$$- 2\pi \int_0^t F_o\,dt - 2\pi \int_0^t f \sin r\,(t - d_2/V)\,dt, \qquad (11)$$

$$\Delta\Theta = \frac{2\pi f}{r}[(\cos rt - 1)(\cos r\,d_2/V - \cos r\,d_1/V)$$
$$+ \sin rt\,(\sin r\,d_2/V - \sin r\,d_1/V)]. \qquad (13)$$

Delete (12)

standing importance have been made in the characteristics of the load-

[1] Presented at the Midwinter Convention of the A. I. E. E., New York, N. Y., February 9, 1926.

[2] "Commercial Loading of Telephone Circuits in the Bell System," B. Gherardi, Trans. A. I. E. E., Vol. 30, 1911, p. 1743.

The Bell System Technical Journal

April, 1926

Development and Application of Loading for Telephone Circuits[1]

By THOMAS SHAW and WILLIAM FONDILLER

SYNOPSIS: A review of the art of loading telephone circuits as practised in the United States. The introductory section briefly reviews the theory of coil loading, and summarizes the principal characteristics of the first commercial standard loading coils and loading systems, thereby serving as a background for the description of the various improvements of outstanding importance which have been made in the loading coils and loading systems during the past fifteen years to meet the new or changing requirements in the rapidly advancing communication art.

These major improvements are described in detail under the appropriate headings (1) Phantom Group Loading, (2) Loading for Repeatered Circuits, (3) Incidental Cables in Open Wire Lines, (4) Cross-talk, (5) Telegraphy over Loaded Telephone Circuits, (6) Loading for Exchange Area Cables, and (7) Submarine Cables. The discussion of these various developments sets forth the relations between the loading features and the associated phases of telephone development, such as the cables, repeaters, telegraph working, and carrier telephone and telegraph systems.

The concluding part of the paper gives some general statistics regarding the extent of the commercial application of loading in the United States, and a brief statement indicative of the large economic importance of loading to the telephone using public.

INTRODUCTION

THE year 1926 marks the fiftieth anniversary of the birth of the telephone, and the completion of the first 25 years of the commercial application of loading to telephone circuits by means of inductance coils inserted at periodic intervals. The present time is thus peculiarly appropriate for a survey of loading developments.

The purpose of this paper is to present a review of the art of loading telephone circuits, as practised in the United States. In a paper[2] presented before the Institute in 1911 Mr. B. Gherardi described the developments in loading up to that time and gave a comprehensive statement of the results obtained. In the present paper, therefore, references to the early developments in loading may be confined to matters that are necessary to the treatment of the subsequent developments in the art.

During the period under consideration many improvements of outstanding importance have been made in the characteristics of the load-

[1] Presented at the Midwinter Convention of the A. I. E. E., New York, N. Y., February 9, 1926.

[2] "Commercial Loading of Telephone Circuits in the Bell System," B. Gherardi, Trans. A. I. E. E., Vol. 30, 1911, p. 1743.

ing coils and in the loading systems, in order to meet new or changing requirements in the rapidly advancing communication art. The more important of these improvements are listed below and will be discussed in the sequence noted:

I. Phantom Group Loading
II. Loading for Repeatered Circuits
III. Incidental Cables in Open Wire Lines
IV. Cross-Talk
V. Telegraphy over Loaded Telephone Circuits
VI. Loading for Exchange Area Cables
VII. Submarine Cables

As a basis for the discussion of the characteristics of commercial loading systems and the various developments which have been made, the elementary theory of loaded lines and a review of the first loading standards will be given. Those interested in the exact mathematical theory are referred to more complete discussions which may be found in the bibliography appended hereto.

Theory[3]. It is convenient to discuss the coil loaded line in terms of its corresponding smooth line, a hypothetical line in which the constants of the inductance coils are assumed to be distributed uniformly along the line.

Table I gives simplified formulas which define the important line characteristics in terms of the primary line constants, the formulas

TABLE I

Approximate Line Formulas

Line Characteristics	Uniform Line Having Zero Inductance	Uniform Line Having Distributed Inductance	
α, Attenuation constant	$\sqrt{\frac{pRC}{2}}$	$\sqrt{\frac{R}{2Lp}} \cdot \sqrt{\frac{pRC}{2}} = \frac{R}{2}\sqrt{\frac{C}{L}}$	(1)
W, velocity of wave propagation	$\sqrt{\frac{2p}{RC}}$	$\sqrt{\frac{R}{2Lp}} \cdot \sqrt{\frac{2p}{RC}} = \sqrt{\frac{1}{CL}}$	(2)
Z_0, characteristic impedance	$\sqrt{\frac{R}{pC}}\ /45°.$	$\sqrt{\frac{Lp}{R}}\ /45° \cdot \sqrt{\frac{R}{pC}}\ /45° = \sqrt{\frac{L}{C}}$	(3)

In the above, α is the real part of the propagation constant; and $W = p/\beta$, in which $p = 2\pi f$ (f = frequency) and β is the wave length constant; *i. e.*, the imaginary part of the propagation constant. The formulas assume the leakage conductance G to be negligibly small; and in the case of the line with inductance, that R is small with reference to $p\,L$; R, L, and C being the line resistance, inductance, and capacitance per unit length.

[3] This section on Theory contains a small amount of discussion not included in the paper as presented.

being so arranged as to indicate directly the nature of the changes which occur when uniformly distributed inductance is added to a uniform line initially having zero inductance.

Inspection of the formulas shows that the addition of distributed inductance:

(a) Reduces the attenuation constant and the velocity, provided that the ratio $R/2L$ is less than p; in practice, this limiting condition is approached only at very low frequencies which usually are of negligible importance in speech transmission.

(b) Increases the impedance, and improves the power factor.

(c) Makes the attenuation, velocity and impedance independent of frequency over the frequency range where R is small with reference to pL; in practice, this condition holds generally, except at the low voice frequencies.

From the standpoint of the power transmission engineer, the general effect of loading in reducing the attenuation losses may be explained in terms of the changes in line impedance noted in (b) above. These impedance changes make it possible for the loaded line to transmit a given amount of power corresponding to speech sounds at a higher line potential and with a (proportionately) lower value of line current than is possible without the loading. In the non-loaded line which is inherently a low impedance line, the series dissipation losses which are proportional to the square of the line current are ordinarily very large relative to the shunt dissipation losses which are proportional to the square of the line potential. Consequently, when the line impedance is increased by a suitable amount, the reduction in series losses is much greater than the increase in shunt losses and a substantial improvement in line efficiency is obtained. The optimum impedance for minimum line losses is that which results in the shunt and series losses being equal. Ordinarily, it is not economical to apply a sufficient amount of loading to reach this condition.

In general, commercial power lines are electrically short in terms of the wave length of the transmitted frequencies and consequently the sending end impedance is very largely influenced by the receiving end impedance. This allows high impedance transmission lines to be obtained by using high ratio transformers at the receiving end to step up the terminal impedance. On the other hand, telephone lines which are of interest from the loading standpoint are electrically long and the sending end impedance is practically unaffected by the terminal impedance. Consequently, the addition of series inductance to the line is the most practical way of increasing the telephone line impedance.

Investigating the question of concentrating the line inductance at uniformly spaced intervals, Professor Pupin gave his famous solution in a paper[4] presented before the Institute in May, 1900. Dr. G. A. Campbell in his paper[5] of March, 1903, also gave a mathematical development of the loading theory along somewhat different lines.

These early investigations showed that a coil loaded line should have several coils per wave length in order to simulate a uniform line. The more closely the coils are spaced the more exact is the degree of equivalence, and when there are ten coils per wave length the equivalence is very close. On the other hand, the cost of the loading increases as the spacing is shortened. Thus, from the standpoint of commercial application, the question "What is the smallest number of coils per wave length that will give satisfactory transmission?" is very important. In the investigation which was made to determine the magnitude of the changes in attenuation, velocity and impedance, as the number of coils per wave length is reduced, abrupt changes in these characteristics were found to occur at the spacing of two coils per actual wave length. The critical frequency at which this spacing applies in a loaded line became known as the cutoff frequency, since at this frequency and higher frequencies the attenuation loss is so extremely large as to amount practically to a suppression, or cut-off effect.

At the cut-off freqeuncy the velocity of the coil loaded line is lower than the velocity of the corresponding smooth line approximately in the ratio of $2:\pi$; consequently, at the cut-off frequency there are approximately π coils per wave length, in terms of the velocity of the corresponding smooth line.

The following expression defines the cut-off frequency in a coil loaded line having zero distributed inductance:

$$f_c = \frac{1}{\pi\sqrt{LsC}} \tag{4}$$

in which

f_c = cut-off frequency,
L = coil inductance,
s = coil spacing,
C = line capacitance per unit length.

[4] "Wave Transmission over Non-Uniform Cables and Long Distance Air Lines," M. I. Pupin, Trans. A. I. E. E., Vol. 17, 1900, p. 445. Refer also to Pupin, U. S. Patents Nos. 652, 230 and 652, 231, June 19, 1900.

[5] "On Loaded Lines in Telephone Transmission," G. A. Campbell, *Philosophical Magazine*, March, 1903.

[If the loaded line has distributed inductance, a correction is required in equation (4).]

The differences between the characteristics of a coil loaded line and its corresponding smooth line are sometimes designated "lumpiness" effects. They are due to repeated internal reflections at the points of electrical discontinuity in the line caused by the insertion of the loading coils. The lumpiness effects are usually small for the frequencies below approximately 75 per cent. of the cut-off frequency. As the frequency exceeds this value, however, the lumpiness effects increase at an accelerated rate.

Figs. 1, 2 and 3 illustrate the differences in the attenuation, velocity, and impedance characteristics of a typical telephone cable, with and without loading. The characteristics of the corresponding smooth loaded line are also given to illustrate the theoretical differences between uniform loading and coil loading. Fig. 1 includes curves

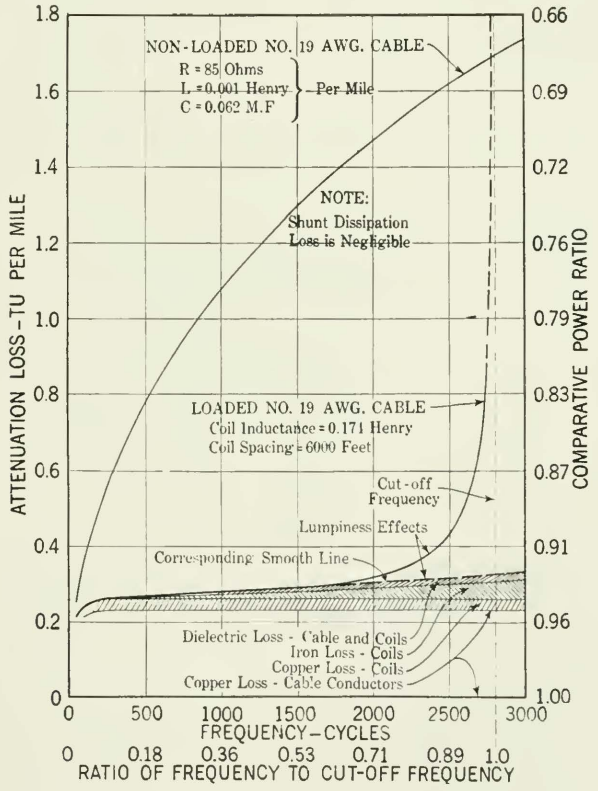


Fig. 1—Attenuation-frequency characteristics of loaded and non-loaded No. 19 A. w. g. cable

which give an analysis of the different types of line losses, (*a*) the "series" losses due to heat dissipation in the conductor and the loading coils, which are proportional to the square of the line current, (*b*) the "shunt" losses due to heat losses in the dielectrics, which are proportional to the square of the line voltage, and (*c*) the lumpiness effects due to internal reflections. The large reduction in the series losses

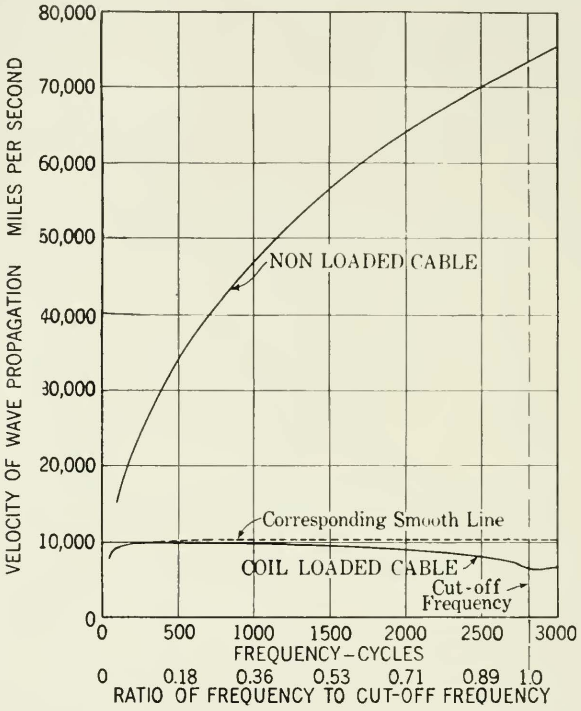


Fig. 2—Velocity-frequency characteristics of loaded and non-loaded No. 19 A. w. g. cables of Fig. 1

accomplished by the loading is clearly indicated in the diagram. A corresponding proportional increase in the shunt dissipation loss also occurs, but as previously noted this effect is small in absolute magnitude relative to the decrease in the series losses. It is interesting to note that the particular type of loading illustrated in Fig. 1 so increases the transmission efficiency of No. 19 A.W.G. cable that the loaded circuit can be used for distances about four times the permissible length of the non-loaded circuits. To obtain this increased transmission range without loading would require wires about eight times as heavy, i.e., No. 10 A.W.G.

Fig. 3 illustrates the dependency of the characteristic impedance of a coil loaded line upon the terminal condition. The most frequently used loading terminations are "mid-section" and "mid-coil." In the mid-section termination, the first loading coil is located at a dis-

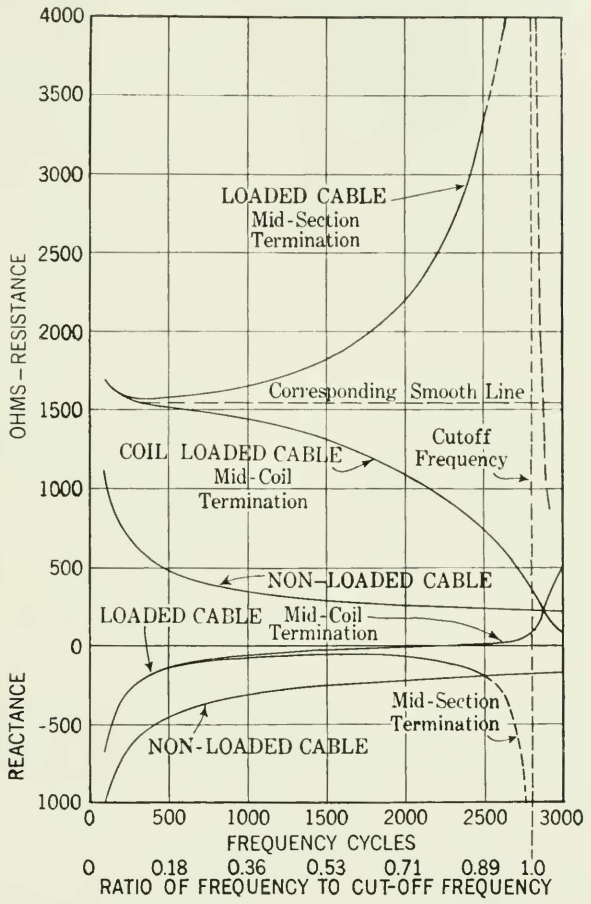


Fig. 3—Impedance-frequency characteristics of loaded and non-loaded No. 19 A. w. g. cables of Fig. 1

tance equivalent to one-half of a regular loading section from the beginning of the line. Mid-coil termination is obtained by installing at the beginning of the line, a coil having one half of the inductance of the regular coils, the first full coil being installed at the end of the first complete loading section. For mid-coil and mid-section terminations, the characteristic impedance is approximately a pure resistance,

which varies with frequency as a complicated function of the ratio of the frequency to the cut-off frequency. With mid-coil termination the impedance-frequency characteristic droops with rising frequency, approaching zero at the cut-off frequency. On the other hand, the mid-section termination has a rising characteristic, approaching infinity at the cut-off frequency.

Early Standard Loading Systems. One of the fundamental questions involved in the early commercial development work was that of determining what range of frequencies should be transmitted in order to furnish a satisfactory grade of speech transmission. The investigation of this point resulted in the adoption of a standard cut-off frequency of about 2,300 cycles. Table II lists the other important transmission characteristics of the first loading systems standardized about 1904 for use on cables:—

TABLE II

First Standard Cable Loading Systems

Loading Designation	Coil Inductance (Henrys)	Coil Spacing (Miles)	Inductance per Mile (Henrys)	Nominal Impedance (Ohms)	Attenuation Loss (TU per mile)		
					19 A.w.g.	16 A.w.g.	14 A.w.g.
Heavy	0.250	1.25	0.200	1800	0.28	0.16	0.11
Medium	0.175	1.75	0.100	1300	0.39	0.21	0.14
Light	0.135	2.5	0.054	900	0.51	0.27	0.17
		(Non-loaded Cable)			1.05	0.74	0.59

NOTE. These data apply to cables having a mutual capacitance of 0.070 μf. per mile and assume loading coils, the electrical characteristics of which are given in Table IV. The nominal impedance is defined by the expression $\sqrt{L/C}$. The new unit of transmission loss (TU) is described in a recent Institute paper.[6]

For open wire loading, only one loading system, known as "Heavy" loading, was standardized. This involved the use of coils having an inductance of 0.265 henry at a spacing of approximately 8 miles. This loading had approximately the same cut-off frequency as the cable loading standards described in Table II. The other important line and transmission characteristics are summarized in Table III.

Loading Coils. The loading coils developed for use in the loading systems described in Tables II and III were of the toroidal type; i.e., they had ring-shaped cores formed by winding up a bundle of insu-

[6] "The Transmission Unit and Telephone Transmission Reference Systems," W. H. Martin, Trans. A. I. E. E., Vol. 43, 1924, p. 797; *Bell System Technical Journal*, July, 1924.

TABLE III

First Standard Open Wire Loading

Wire Diameter (In.)	Loading Condition	Constants per Loop Mile			Nominal Impedance Ohms	Attenuation Loss TU per Mile
		R Ohms	*L* Henrys	*C* Mf.		
0.104	Non-loaded	10.4	0.0037	0.0084	660	0.075
0.104	Loaded	11.1	0.037	0.0086	2100	0.031
0.165	Non-loaded	4.14	0.0034	0.0091	610	0.033
0.165	Loaded	4.8	0.037	0.0094	2000	0.014

NOTE. Transmission efficiency figures assume dry weather insulation conditions, 5 megohm-miles, or better.

lated fine wires on a suitably shaped spool. The core wire was 38 A. w. g. (0.004 in. diameter).

The wire used in the cable loading coil cores was a commercial grade of mild steel, hard drawn under conditions which gave it an initial permeability of 95. The term "initial permeability" signifies the permeability at very weak magnetizing forces; i.e., below 0.1 gilbert per cm. The core wire used in the open wire loading coils was drawn from the same stock, but differences in the drawing and annealing treatments gave it an initial permeability of about 65. This core wire had lower eddy current and hysteresis losses than the 95-permeability wire. A black enamel insulation was used on the 95-permeability wire. A celluloid-shellac compound which could be applied at a lower temperature was used on the 65-permeability wire.

As illustrating the magnitudes involved, it may be noted that in order to meet the service requirements, the coils were designed so that for telephone currents of the order of 0.001 ampere, the magnetizing force H has a value of about 0.04 gilbert per cm., corresponding to a flux density of approximately $B=2$ gauss.

The winding space on the cores was divided in half by means of fiber washers, and the winding was applied in two equal sections, one being located on each half of the core. In installing the coils, one of these windings was inserted in one line wire and the other winding in the other line wire, so connected that the mutual inductance between windings aided the self-inductance for current flowing around the circuit through both windings.

The high costs of the open wire lines warranted considerable refinement in the design of the open wire coils. They were, therefore, made much more efficient and correspondingly larger than the cable coils. They were wound with insulated stranded wire and had much

lower core losses. Another important difference between the open wire and cable coils was the use of high dielectric strength insulation in the open wire coils. The coils were subjected to a breakdown test at 8,000 volts (effective a-c.) and were protected in service by means of a special type of lightning arrester having non-arcing metal electrodes designed to operate at 3,500 volts direct current.

Table IV lists the principal characteristics of the loading coils

TABLE IV

First Standard Loading Coils

Type Loading	Coil Code No.	Inductance (Henrys)	Average Resistance		Overall Dimensions	
			D-C. (Ohms)	1000-Cycle (Ohms)	Diameter (In.)	Height (In.)
Open Wire	501	0.265	2.5	5.9	9	4
Cable	506	0.250	6.4	22.3	4 1/8	3 1/4
"	508	0.175	4.2	13.0	"	"
"	507	0.135	3.2	9.1	"	"

NOTE. Effective resistance values apply for a line current of 0.002 ampere.

initially used in the standard loading systems listed in Tables II and III.

Loading Coil Cases. The cases used for potting the cable loading coils were designed so that they could be installed in underground manholes or on pole fixtures.

The general method of assembly is to dry the loading coils thoroughly and then impregnate them under vacuum with a moisture-proofing compound. The coils are then mounted on wooden spindles, adjacent coils being separated by iron washers. After carefully adjusting the individual coils to meet the electrical requirements, the spindles of coils are cabled to a short length of lead-covered cable which is referred to as a "stub" cable. Cast-iron cases with iron partitions were designed so as to provide a shielded compartment for each spindle of coils.

Commercially manufactured toroidal coils may have small irregularities in their windings resulting in a weak stray field which tends to cause cross-talk. The iron washers between coils and the partitions between spindle groups of coils provide effective cross-talk shields.

After placing the spindles of coils in the various compartments, the case is filled with a moisture proofing compound. The lead-sheathed cable stub is brought through a brass nipple in the cast iron cover of the case, and the cover is then bolted to the case. By means of a special design of case and cover joint, a double seal is provided to prevent entrance of moisture at this point. A wiped joint is made between the lead sheath of the cable and the brass nipple.

The conductors in the stub cable have an appropriate color scheme in their insulation to identify the terminals of the loading coils, thus facilitating splicing of the coils into the line circuits. A series of multi-spindle cases was standardized, ranging in capacity from 21 to 98 coils. Smaller quantities of coils were potted in a single spindle pipe type case.

Generally similar assembly and potting methods were used for the open wire coils, the important differences being first, that the open wire coils were always mounted in individual cases designed for mounting on pole fixtures, and secondly, that the coil terminals were brought out of the case in individual rubber-insulated leads.

I. Phantom Group Loading

In Mr. Gherardi's paper reference was made to the development of means for (*a*) phantoming loaded circuits and (*b*) loading phantom circuits. The large plant economies made possible by these developments have resulted in extensive applications of these principles.

The following discussion will consider first the coil winding schemes, after which the transmission characteristics of the loading systems and the electrical characteristics of the loading coils will be briefly described.

Loading Methods. Fig. 4 schematically illustrates the Bell System standard method for loading phantom circuits and side circuits of phantoms.[7]

The loading problem is to introduce the desired inductance into each of the three circuits of a phantom group without causing objectionable unbalances. The method illustrated in Fig. 4 involves individual loading coils for each circuit, the design being such that the side circuit coils are substantially non-inductive to the phantom circuit, while the phantom loading coil is substantially non-inductive to the side circuits. These desirable results require close magnetic coupling between the line windings in each coil. Consequently, in

[7] U.S. Patents No. 980,021 "Loaded Phantom Circuit," G. A. Campbell and T. Shaw. No. 981,015 "Phantomed Loaded Circuit," T. Shaw.

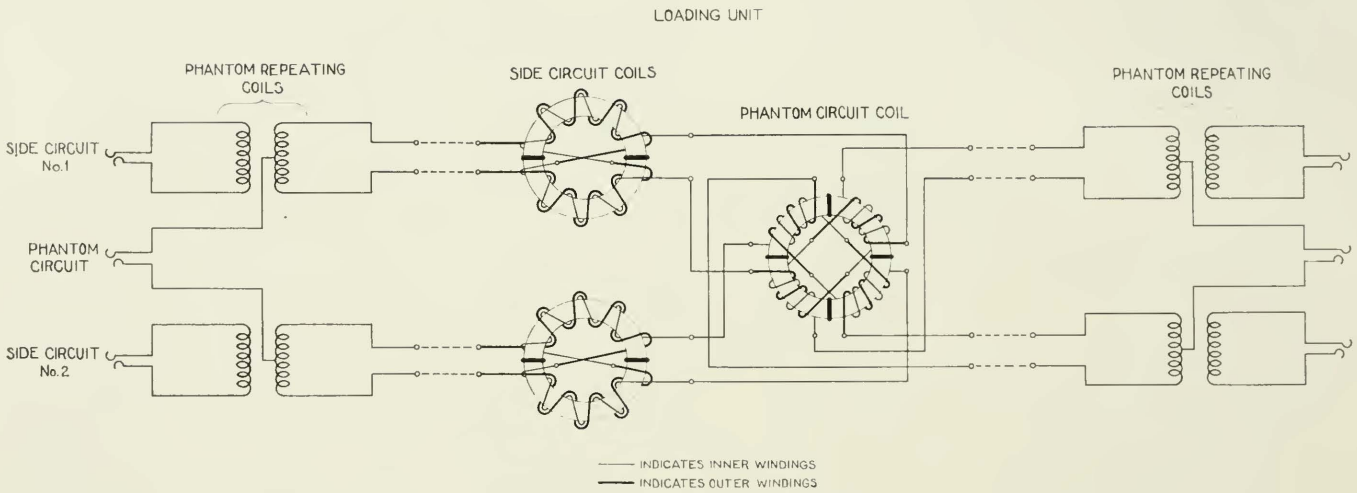


Fig. 4—Bell system standard method of loading phantom circuits and their side circuits

the side circuit coils each line winding is, in effect, distributed evenly about the entire core. The necessary high degree of symmetry required by balance considerations is obtained by dividing each line winding into two equal sections and interleaving them with the sections of the other line winding; thus each complete line winding consists of an inner section winding on one-half of the core and an outer section winding on the opposite half core. Similar design principles are applied to the phantom loading coils, with added complications, however, arising from the increased number of line windings. Each of the four line windings consists of an inner section winding located on one core quadrant and an outer section winding located on the opposite core quadrant, the two line windings associated with a given side circuit being distributed about the same pair of opposite core quadrants. In arranging the windings on the core, precautions are taken to secure a symmetrical arrangement of the direct admittances among the line windings and from the line windings to the core and the case.

It is interesting to note that the three-coil loading scheme illustrated in Fig. 4 was employed in the Boston-Neponset cable, installed in 1910, which was the first successful installation of loaded phantom circuits in the world. Other schemes of phantom group loading using two-coil and four-coil arrangements have been developed here and abroad, but none of them is considered to be as satisfactory as the scheme illustrated in Fig. 4 from the standpoint of service and cost. These other schemes are described in a recent article [8] which compares them with the scheme above described.

Loading Systems. In adapting the circuits to phantom working, the electrical constants of the two-wire circuits were changed as little as possible in making them suitable for use as side circuits of phantoms. In the cables, two pairs having different lengths of twist were twisted into quad formation on a still different length of twist. The required balance was obtained on open wire lines by cutting in a large number of additional transpositions.

The construction methods chosen resulted in the phantom circuits having approximately 60 per cent. greater distributed capacity than their side circuits, and a lower distributed inductance, approximately in inverse proportion. It was obviously desirable to install the phantom circuit coils at the same points as the side circuit coils; accordingly, in working to the same standard of cut-off frequency, the relative circuit constants summarized above resulted in the phantom

[8] "Commercial Loading of Telephone Cable," W. Fondiller, Electrical Communication, July 1925.

loading systems having a nominal impedance approximately 60 per cent. as high as their associated side circuit loading systems. The transmission efficiency of the phantom circuit was 20 to 25 per cent. better than that of its associated side circuits, on which basis, the phantom circuits were suitable for use over somewhat longer distances than their side circuits.

Cable Loading. Data regarding the general characteristics of the first phantom group loading systems standardized for use on quadded telephone cables are given in the first four rows of Table V. These loading systems were used principally on interurban toll cables. Because of the extra cost of the terminal and signaling equipment, and other factors involved in phantom working, it was not economical to use phantom circuits in the shorter lengths of loaded cable ordinarily involved in exchange area connections.

As soon as the development work on quadded toll cables and phantom group loading had progressed to a point where satisfactory commercial results were assured, active development work commenced on the Boston-New York-Washington cable project, involving the use of coarse gage quadded conductors and new types of high efficiency loading coils designed especially for use on the coarse gage wires. The Boston-Washington cable was the first link in a rapidly growing network of toll cables which now interconnects the large population centers of the Atlantic Seaboard and the upper Mississippi Valley region, providing increased reliability of service as compared with open wire lines.

It should be kept in mind that at the time under discussion (1910–1911) no commercially satisfactory type of telephone repeater was available. Accordingly, in order to assure satisfactory service between Boston, New York, Washington, and intermediate points, it was necessary to provide 10-A.w.g. and 13-A.w.g. conductors in the new cable. Cost studies showed it to be desirable to use a special weight of loading intermediate between the old heavy and medium loading systems, which was therefore designated "Medium-heavy" loading. Information regarding this special loading is given in Items 1 and 2 of Table V. In items 3 and 4, corresponding data are given on the "high-efficiency" heavy loading designed for coarse gage conductors. This heavy loading was used on certain sections of the Boston-Washington cable where plant construction reasons made it desirable to install the coils in existing loading manholes installed at heavy loading spacing.

From the last column of Table V it is seen that there is very little difference between the efficiencies of the heavy and the medium-

TABLE V

First Loading Standards for Quadded Toll Cables

Item	Loading Designation	Type Circuit	Coil Inductance (Henrys)	Coil Spacing (Miles)	Nominal Impedance (Ohms)	Attenuation Loss—TU per mile			
						19 A. w. g.	16 A. w. g.	13 A. w. g.	10 A. w. g.
1	Medium-Heavy	Side	0.210	1.4	1500			0.085	0.050
2	" "	Phantom	0.130	1.4	950			0.069	0.040
3	Heavy	Side	0.250	1.25	1850			0.081	0.050
4	"	Phantom	0.155	1.25	1150			0.066	0.042
5	Heavy	Side	0.250	1.25	1850	0.24	0.14		
6	"	Phantom	0.155	1.25	1150	0.20	0.12		
7	Medium	Side	0.175	1.75	1300	0.31	0.17		
8	"	Phantom	0.106	1.75	800	0.26	0.14		

NOTES: A capacitance of 0.062 μf. per mile is assumed in side circuits and 0.100 μf. per mile in the phantom circuit. The pair capacitance value is smaller than that assumed in Table II, due to improvement in the cables.

All of the above loading systems have a cut-off frequency of about 2300 cycles.

heavy loading systems when used on 10-A.w.g. conductors. This explains the more general use of the medium-heavy loading, which was less expensive because of the greater distances between coils. The effects under discussion are due to the part played by the loading coil resistance. The loading coils themselves conformed as closely as practicable to the cost-equilibrium principle:—a condition of cost balance where a small improvement in transmission would require approximately equal expenditure whether by improving the loading or by adding copper to the cable conductors. On this basis, a somewhat less expensive grade of coil was used on the 13-A.w.g. wires than on the 10-A.w.g. wires. The grade of coils developed primarily for use on 16 and 19-A.w.g. cables, giving transmission results illustrated in Items 5-8 of Table V, was in turn less expensive than the "high efficiency" coils. In each case, since the phantom circuits were somewhat more efficient than their associated side circuits, a somewhat higher grade coil was used in the phantom circuits than in the side circuits.

Open Wire Phantom Loading. Phantom loading came into general use on open wire lines at about the same time as on quadded cables. In general, the methods used in applying phantom group loading to the open wire lines were used for the cable systems. The line characteristics for the side circuits were practically the same as for the original non-phantomed circuits (Table III); the principal difference being that caused by the small resistance of the phantom loading coils. The important linear and transmission characteristics of the phantom circuits are given in Table VI. The phantom loading coil had an inductance value of 0.163 henry.

Loading Coils. Table VII gives general information regarding the first standard side circuit and phantom loading coils used in the phantom group loading systems listed in Tables V and VI. The coils

TABLE VI

First Standard Open Wire Phantom Loading

Wire Diameter (In.)	Loading Condition	Constants per Loop Mile at 1000 Cycles			Nominal Impedance (Ohms)	Attenuation Loss TU per Mile
		R (Ohms)	*L* (Henrys)	*C* (Mf.)		
0.104	Non-loaded	5.2	0.0022	0.0141	400	0.064
0.104	Loaded	5.8	0.023	0.0141	1300	0.027
0.165	Non-loaded	2.1	0.0021	0.0154	400	0.028
0.165	Loaded	2.6	0.023	0.0154	1200	0.012

TABLE VII

First Standard Loading Coils for Phantom Working

Type Line	Inductance (Henrys)	Coil Code No.	Type Circuit	Average Resistance-Ohms		Overall Dimensions	
				D-C.	1000 cycles	Diameter (In.)	Height (In.)
Open-Wire	0.265	512	Side	5.0	8.4	9.0	4.0
	0.163	511	Phantom	2.5	4.4	11.0	4.9
10-A. w. g. Cable	0.210	520	Side	3.8	6.6	8.5	3.5
	0.130	519	Phantom	1.9	3.4	10.4	4.0
	0.250	532	Side	4.1	7.8	8.5	3.5
	0.155	531	Phantom	2.1	3.9	10.4	4.0
13-A. w. g. Cable	0.205	538	Side	6.0	9.2	5.7	2.5
	0.130	521	Phantom	3.0	4.5	7.9	3.0
	0.250	534	Side	6.6	10.7	5.7	2.5
	0.155	533	Phantom	3.3	5.3	7.9	3.0
16 and 19-A. w. g. Cable	0.250	515	Side	8.9	23.1	4.6	2.4
	0.155	530	Phantom	4.4	11.9	5.9	2.9
	0.175	514	Side	5.4	14.4	4.6	2.4
	0.106	513	Phantom	2.7	7.1	5.9	2.9

NOTE. The resistance data apply to circuits of a complete phantom group; *i.e.*, the side circuit data include effects of the phantom coils, and phantom circuit data include effects of the side circuit coils. Effective resistance values correspond to line current of 0.002 ampere.

designed for open wire lines and for 10-A.w.g. cable had 65-permeability wire cores and stranded copper windings. The coils designed for 13-A.w.g. cables had 65-permeability wire cores and non-stranded copper windings. The other coils had 95-permeability wire cores.

Potting Features. The general practise for cable loading is to pot side circuit and phantom loading coils in the same case as phantom groups, since this has important installation and transmission advantages. The phantom coils, being considerably larger than the side circuit coils, are mounted in separate spindle compartments. The cross-connections between the side circuit and phantom coils are made within the case, in order to reduce the amount of splicing required in the field. Thus, the stub cable contains only the conductors to be spliced to the "east" and "west" conductors in the line cable. Quadded construction is used in the stub cable of all loading coil cases for phantom loading in order to avoid serious capacitance unbalances.

The multi-spindle cases used in potting the small size coils for 16 and 19-A.w.g. cables ranged in capacity from 12 to 24 phantom units. The larger size coils used on the coarser gage cables were potted in smaller complements.

Occasionally it is desirable to install side circuit loading alone and to install the phantom loading at a later period. Accordingly, cable loading coil cases were designed to meet these conditions. The open wire coils were potted in individual cases.

II. Loading for Repeatered Circuits

General. The development of telephone repeaters to the point where they could be used for commercial service in extending the range of telephone transmission was the beginning of a new era in the communication art. In this development work, the adaptation of the lines to the requirements of repeater operation was secondary in importance only to the development of satisfactory repeater elements and circuits for associating the repeater elements with the line. The reader is referred to an Institute paper by Messrs. B. Gherardi and F. B. Jewett [9] for general information regarding telephone repeaters and to a more recent Institute paper by Mr. A. B. Clark[10] for a general discussion of subsequent developments in the application of repeaters to long telephone circuits.

The early work on the line problem was primarily concerned with obtaining a sufficiently high degree of regularity in the line impedance-frequency characteristics, so that the requisite high degree of balance could be obtained and maintained between the line and the repeater balancing network. Later on, particularly in preparing for the application of telephone repeaters to long toll cables, such as the New York-Pittsburgh-Chicago cable, it became necessary to change the fundamental transmission characteristics of the loading.

Early Work—Reduction of Line Irregularities. Commercial telephony, requiring two-way transmission, imposes severe balance requirements on repeater circuits over the entire band of frequencies which the repeater is designed to transmit, in order to avoid singing or distortion due to near singing. Within certain limitations, the higher the degree of balance between the line and the balancing network circuit, the higher will be the permissible amplification gain of the repeater.

[9] "Telephone Repeaters," B. Gherardi and F. B. Jewett, Trans. A. I. E. E., Vol. 38, 1919, p. 1287.

[10] "Telephone Transmission over Long Cable Circuits," A. B. Clark, Trans. A. I. E. E., Vol. 42, 1923, p. 86; *Bell System Technical Journal*, Jan., 1923.

The practical solution of this fundamental repeater-line balance problem required (*a*) the construction of lines having extremely regular impedance characteristics over the frequency band which the repeater is designed to transmit and (*b*) the development of balancing networks [11] capable of accurately simulating the sending-end impedance characteristics of the improved lines throughout this frequency range. On account of the great difficulty of getting a high degree of balance at frequencies near the cut-off frequency of the loading, partly due to line irregularity effects and partly due to network design complications, it has been found desirable to use electric wave filters [12] in the repeater sets which cut off at a frequency below the cut-off frequency of the loading. This margin of cut-off effects is usually 200 cycles or more, depending upon the repeater design and the type of loading involved.

The "regular" line referred to in (*a*) is one which is free from impedance irregularities. In the case of loaded lines, the loading coils should have very closely the same inductance values, and the sections of line between loading coils should have closely the same value of capacitance. These uniformity features should be permanent, which requires that the coils should have a high degree of stability in their inductance characteristics; i.e., they should be capable of resisting the magnetizing effects of abnormal service conditions. Some of the older types of coils did not meet this requirement. The satisfactory way in which these fundamental coil requirements are fulfilled in the newer types of coils will be described in a subsequent section.

Uniformity in the loading section capacitance values involves uniformity in cable and line capacitance values as well as precision in the coil spacing. In toll cable loading the maximum deviations from the average spacing are kept below 2 per cent., and the average deviations are in the order of 0.5 per cent. or less.

In exceptional cases where physical obstructions are encountered in reducing the spacing deviations to a sufficiently low value, use is made of "building-out condensers" or "building-out stub cables" to normalize the capacitance of loading sections.[13] Abnormally long loading sections can usually be split up into two sections, one or both of which may then be "built out" to the nominal standard capacitance values.

Transcontinental Lines—High Stability Loading Coils. The in-

[11] R. S. Hoyt "Impedance of Loaded Lines and Design of Simulating and Compensating Networks," *Bell System Technical Journal*, July 1924.

[12] U. S. Patents Nos. 1,227,113, and 1,227,114—G. A. Campbell.

[13] U. S. Patent No. 1,219,760—John Mills and R. S. Hoyt.

auguration of commercial transcontinental telephone service over the New York-San Francisco line in January, 1915, marked the first commercial application of these general improvements in regularity of line construction, including the use of an improved type of loading coil.

In the extensive field work which was done in preparing for transcontinental telephone service, it was found that the inductance values of a considerable percentage of the open-wire loading coils then in use (Nos. 511 and 512 types, Table VII) had changed appreciably from the nominal values to which they were adjusted at the factory prior to shipment, and that these changes were due to core magnetization caused by abnormal currents induced by lightning discharges. In some cases abnormal currents induced by power transmission lines or electric railway distribution systems were responsible for the loading coil magnetization.

The inductance changes were not sufficiently large to have serious reactions on transmission over non-repeatered circuits. Although individual coils varied in inductance from time to time, the general average of groups of coils was fairly constant. The effects of these individual variations on the impedance of the line were, however, too large to permit satisfactory operation with telephone repeaters. Some experiments made with improved lightning arresters, in an effort to reduce the coil magnetization trouble, were unsuccessful.

The solution of the problem of repeatering loaded open-wire circuits required the development of loading coils which would be stable magnetically when subjected to extreme conditions of magnetizing current in the windings. The requirement was laid down for these coils that the inductance to speech currents should not be affected more than about 2 per cent. when a magnetizing current of two amperes was passed through either line winding. In view of the fact that the extreme residual magnetizing effect of this current on the No. 511 and No. 512 loading coils was approximately 30 per cent., it will be appreciated that this imposed a very severe stability requirement.

The design adopted involved the use of air-gaps in the cores of the iron wire core loading coils.[14] Two air-gaps were employed at opposite points in the cores and suitable clamping means were provided to hold the coil halves in proper alinement. The use of only two air-gaps in the cores of the phantom loading coil brought in unbalance tendencies not present in older designs, which were corrected by special refinements in the design.

The use of a magnetic circuit having "ends," while effective for producing self-demagnetization, brought in troublesome magnetic

[14] U. S. Patents Nos. 1,289,941 and 1,433,305—Shaw and Fondiller.

leakage which necessitated special potting methods. Because of the economy of cast-iron loading coil cases, it was decided to continue their use, but to increase their dimensions sufficiently to reduce eddy-current losses in the case to a tolerable point.

The air-gap type loading coils designed for the transcontinental circuits, coded Nos. 549 and 550 for the phantom and side circuits respectively, were more generally potted as phantom loading units than as individual coils, and in such instances the cross-connections between the phantom and side circuit coils were made inside the case. Important advantages of this arrangement were that the leakage losses during periods of low line insulation were greatly reduced as well as the liability of wrong connections of windings during the installation work. Fig. 5 is a photograph of an installation of open



Fig. 5—Typical open wire loading installation
Showing four phantom group (3-coil) cases and nine individual coil cases

wire loading coils illustrating both the individual coil and loading unit methods of potting.

Table VIII contains data on the air-gap coils standardized for open-wire circuits. It will be noted that these coils are somewhat less efficient from the standpoint of effective resistance than the older type coils (Nos. 511 and 512) listed in Table VII, though having

TABLE VIII

High Stability Coils Having Wire Cores with Air Gaps

Type Loading	Coil Code No.	Type Circuit	Inductance	Average Resistance Ohms		Overall Dimension Inches	
			Henrys	D-C.	1000-Cycles	Diameter	Hgt.
Open Wire	550	Side	0.245	5.4	11.1	8.1	3.9
" "	549	Phantom	0.150	2.7	6.4	10.0	4.0
10 and 13	556	Side	0.248	7.0	14.0	5.6	2.9
A. w. g. Cable	555	Phantom	0.154	3.5	7.0	7.5	3.6
10 and 13	558	Side	0.200	6.2	10.9	5.6	2.9
A. w. g. Cable	557	Phantom	0.135	3.1	5.9	7.5	3.6

NOTES. Open-wire coils used in Loading Systems, Tables III and VI. Cable coils used in Loading Systems, Table V.

Resistance data apply to side circuits and phantom circuits of complete phantom groups. Effective resistance values are for 0.002 ampere line current.

marked superiority over the latter with regard to magnetic stability. To assist in getting maximum line regularity, the Nos. 549 and 550 coils were adjusted in the factory to meet ± 1 per cent. inductance precision limits. In the older types of coils ± 5 per cent. deviations had been allowed. The nominal inductance values of the Nos. 549 and 550 coils are somewhat below those of the Nos. 511 and 512 coils, the inductance difference corresponding roughly to the average magnetization effect of normal service conditions on the older types of coils.

The solution of the transcontinental line problem involved improvements in the regularity of the coil spacing as well as improvements in the magnetic stability of the coils. The line "clearing up" work usually involved a great deal of retransposing, since cross-talk considerations made it necessary to have the coils placed at balanced or neutral points in the transposition layout.

In the case of coarse gage cable circuits, such as the Boston-Washington and other toll cables installed prior to the advent of repeaters,

the new requirements were met by the design of an air-gap type of wire-core coil on which data are given in Table VIII. They were somewhat smaller and not quite so expensive as the improved open-wire coils.

Compressed Powdered Iron Core Loading Coils. It soon became evident that the economical extension of the toll plant would involve the general introduction of telephone repeaters in cable as well as open-wire circuits. The use of telephone repeaters made it possible to supersede the coarse gage conductors by 16 and 19-A.w.g. conductors for toll connections, and this greatly increased the need for an efficient and stable loading coil of lower cost than the air-gap wire core coil.

As a result of investigations carried on over a period of several years, there was developed for commercial use early in 1916 a new magnetic material, compressed powdered iron, which has been of the utmost value in loading coil design.[15] This improved magnetic material is described in a paper presented before the Institute by B. Speed and G. W. Elmen [16] which also discusses the electrical and magnetic properties of the material.

Briefly, the method of production consists of grinding electrolytically deposited iron to the desired fineness, insulating the particles of iron, and finally compressing these insulated particles in steel dies at such very high pressures as to consolidate the mass into a ring, the specific gravity of which is substantially equal to that of solid iron. The rings are then stacked in a manner similar to laminations of sheet material to form a core of the desired dimensions. Though the separate rings are approximately 0.2 in. thick, the insulation between the individual particles is so effective that despite the use of molding pressures of 200,000 lb. per sq. in., the eddy current loss in a powdered iron core is less than that obtainable with 0.004 in. iron wire. Depending on the heat treatment and the amount of insulation, the initial permeability can be varied from approximately 25 to about 75. The specific resistance is about 20,000 times that of ordinary iron. The permeability can be controlled within comparatively narrow limits by the manufacturing processes, thus making for greater uniformity. The great advantage of this material for loading coils, however, lies in its self-demagnetizing property. The powdered iron core by virtue of its very numerous, though extremely small dis-

[15] U. S. Patents No. 1,274,952, B. Speed; 1,286,965, G. W. Elmen; 1,292,206, J. C. Woodruff.

[16] "Magnetic Properties of Compressed Powdered Iron," B. Speed and G. W. Elmen, Trans. A. I. E. E., Vol. 40, 1921, p. 1321.

tributed air-gaps, affords a means for constructing magnetically stable cores without the production of poles and their attendant magnetic leakage.

Fig. 6 gives photographs of a standard compressed iron powder core ring such as is used in the cores of toll cable loading coils; a

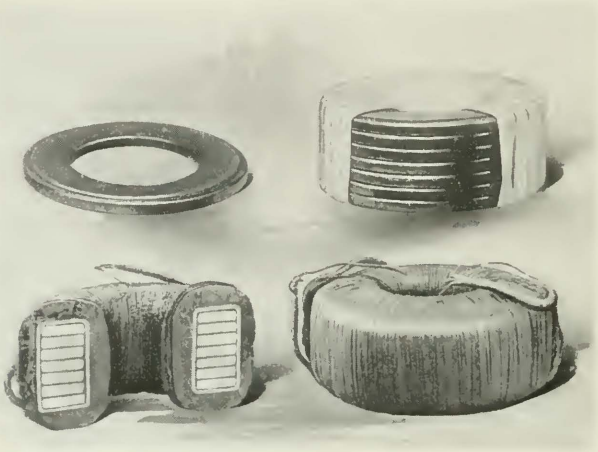


Fig. 6—Compressed powdered iron core loading coil

completely assembled core with part of the core taping removed; a completely wound coil of the side circuit type; and a coil in cross-section. Table IX gives general data regarding typical coils.

The first application of powdered iron cores was to replace some of the 95-permeability wire core loading coils in 16 and 19-A.w.g. cable. The effective permeability of the 95-permeability wire cores, making correction for air spaces and insulation, was approximately 60, and accordingly, the replacing powdered iron cores were designed to have the same effective permeability.

As a result of further developments in the direction of applying vacuum tube repeaters to loaded cable circuits, it became necessary with the extension of the length of these circuits to improve the characteristics of the loading coils. This led to the development of an improved grade of powdered iron core having an initial permeability of 35 which corresponds closely to the effective permeability of cores using iron wire having a permeability of 65. It was decided that for circuits such as interoffice trunks and short cables which would not be operated with superposed telegraph, the 60-permeability compressed iron core coils should be used; while for toll cable work

involving repeatered composited circuits, 35-permeability cores should be employed. All of the compressed powder core coils intended for repeatered circuits were adjusted to meet ± 2 per cent. inductance limits.

The effective resistance-frequency characteristics of 95-permeability and 65-permeability wire core coils and 60-permeability and 35-permeability powdered iron core coils having the same inductance (0.174 henry) and the same over-all sizes are given in Fig. 7. The large improvement as to freedom from residual magnetization effects afforded by the 35-permeability powdered iron core, compared with the 65-permeability wire core is evident from the curves of Fig. 8. The effective resistance and inductance variation with current strength are shown in Fig. 9 for a 35-permeability powdered iron core coil. The remarkable property of these cores of maintaining constancy of permeability is shown by the change of only 1 per cent. in permeability as the current strength varies 400 per cent. from, say 0.001 to 0.005 ampere.

It is interesting to note that after the process had been fully worked out and production was running on a commercial scale, the cost of the improved cores was comparable with that of the wire cores which they replaced.

TABLE IX

Typical Compressed Powdered Iron Core Loading Coils

Coil Code No.	Core Permeability	Inductance (Henrys)	Type Circuit	Resistance Ohms		Dimensions Inches	
				D-C.	1000-Cycles	Diameter	Height
562	60	0.245	Side	11.4	25.8	4.5	2.1
561	60	0.155	Phantom	5.7	11.7	6.3	3.0
564	60	0.174	Side	6.6	15.4	4.5	2.1
563	60	0.106	Phantom	3.3	6.7	6.3	3.0
582	35	0.245	Side	15.9	21.8	4.7	2.4
581	35	0.155	Phantom	8.0	10.0	6.7	3.1
584	35	0.174	Side	10.8	14.1	4.7	2.4
583	35	0.106	Phantom	5.4	6.6	6.7	3.1
584	35	0.174	Side	12.1	15.3	4.7	2.4
587	35	0.063	Phantom	6.1	7.0	4.7	2.8
590	35	0.044	Side	4.0	4.6	4.7	2.4
591	35	0.025	Phantom	2.0	2.0	4.7	2.8

NOTE. Resistance values apply to side circuits and phantom circuits of complete phantom groups. Effective resistance corresponds to 0.002-ampere line current.

These coils are used in the loading systems listed in Tables V and X.

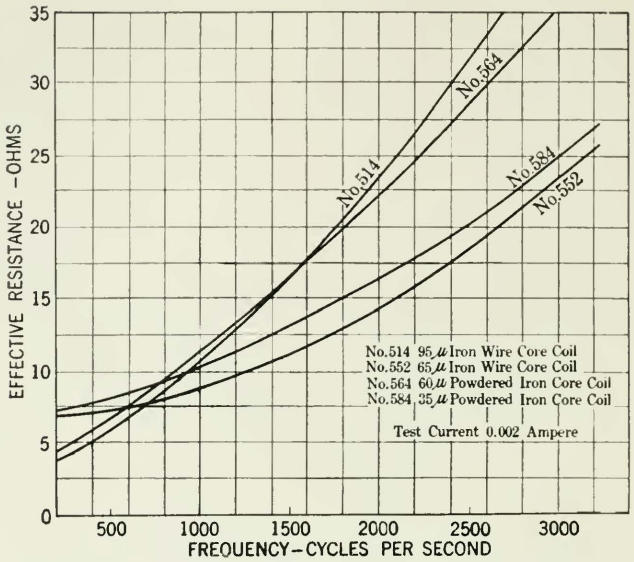


Fig. 7—Effective resistance-frequency characteristics toll cable loading coils

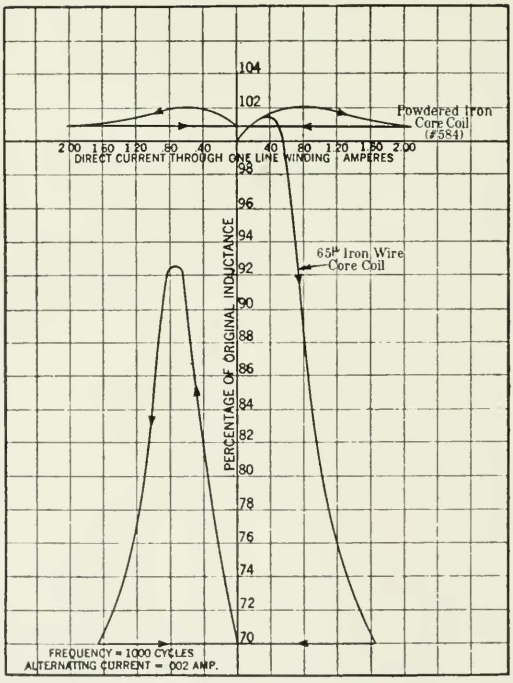


Fig. 8—Residual magnetization characteristics of compressed powdered iron core and iron wire core loading coils

In connection with the development of the new core material which was undertaken as a part of the loading coil development program, an enormous amount of work was involved which would not ordinarily be associated with loading coil design work. For instance, there were undertaken chemical studies on electro-deposition of iron

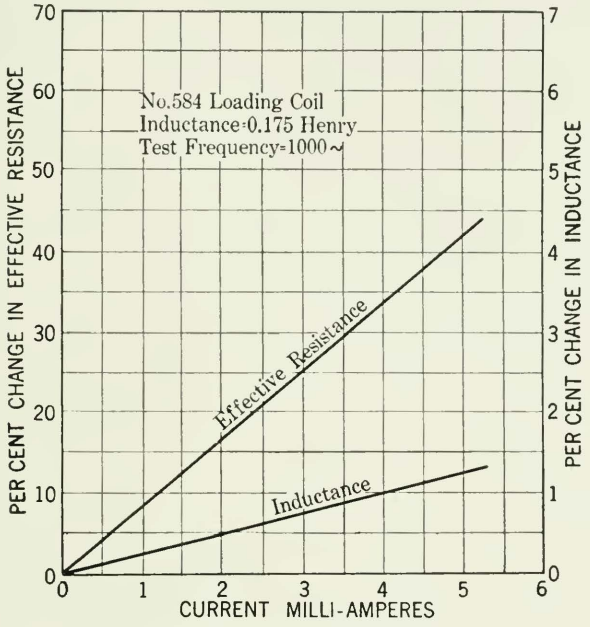


Fig. 9—Variation of Inductance and Effective resistance
With line current in 35-permeability compressed powdered iron core loading coil

and methods of insulating the iron particles, metallurgical studies of the production of finely divided iron by various means, refinements in shielded electrical measuring equipment for accurate determination of small core losses at voice frequencies, development of special permeameters to make possible the rapid determination of the permeability of rings, the design of the steel moulding dies, selection of suitable grades of alloy steel to withstand the enormous pressures, and also various other special problems. These are mentioned here as illustrative of the scope of the problem of developing this new core material.

It is of interest to note that the compressed powdered iron core loading coil has been adopted also as the international standard in Europe for repeatered circuits.[17]

[17] Minutes of Second Conference of Permanent Commission, Le Comité Consultatif Internationale de Communications Telephonique a Grande Distance, page 55—p. 119, English Version.

New Requirements for Cable Loading Systems. In the first commercial applications of telephone repeaters, the new features in the loading were the improved types of coils already described and the improved precision of spacing the coils. No fundamental changes were made in the loading systems then standard.

The completion of the development of a satisfactory commercial type of telephone repeater marked the beginning of a long period of experimental work for the purpose of determining the commercial possibilities of the use of repeaters over long cable circuits. When loaded cables of improved impedance regularity became available, long circuits were built up for experimental purposes by looping back and forth. As the length of these circuits was increased, phenomena not previously observed in cable circuits became increasingly troublesome, and it became apparent that it would be necessary to develop new loading systems having improved velocity and higher cut-off frequency characteristics in order to realize the full possibilities of repeaters in extending the range and reducing the cost of long distance telephone service over cables.

The disturbances above mentioned were found to be due to:

(*a*) Echo effects.
(*b*) Velocity distortion.

These phenomena originate in the lines themselves and are made more apparent by the amplifying action of the repeaters. They are present in non-repeatered circuits but not to a noticeable degree. It is the combination of the extreme length of the circuit and the use of repeaters to keep the over-all loss low that makes the disturbances troublesome.

Echoes. Echoes are due to unbalance currents; i.e., to the reflection of electrical energy at points of impedance irregularity in the circuits. When the circuit is so long that the time of transmission from the point of reflection to the disturbed subscriber is appreciable, there will be echo effects unless the losses in the circuit are so large as to cause the reflected energy to become inappreciably small. On such circuits it may be necessary to work the repeaters at gains well below those at which "singing" occurs or distortion due to "near singing" is experienced.

Since the time of transmission is such an important factor in echo phenomena, reductions in the harmful effects of these disturbances have been obtained in the improved loading systems which have been developed for use on long repeatered circuits, by substantially increasing the velocity of transmission. Recently there has become

commercially available a device known as an "echo suppressor" which interrupts the path of the echoes without disturbing the main transmission. A description of the device and its field of application was given in a recent Institute paper.[18]

Velocity Distortion. In a coil loaded line the steady state velocity of wave propagation varies with frequency. At the upper frequencies the velocity change is principally due to lumpiness effects of the loading and is, therefore, a function of the ratio of the frequency under consideration to the cut-off frequency. As illustrated in Fig. 2, the departure of the actual velocity from the nominal velocity of the corresponding smooth line ($\sqrt{1/CL}$) increases as the frequency is raised, the rate of change increasing rapidly as the cut-off frequency is approached. At frequencies below approximately 0.3 of the cut-off frequency the coil loaded line has substantially the same velocity characteristics as the corresponding smooth line; when the frequency is further reduced, the departure of the actual velocity from the nominal velocity increases as a function of the ratio of the line resistance to the inductive reactance per unit length.

As a result of these velocity-frequency relations, a long loaded repeatered circuit may have seriously objectionable quality, even when the attenuation-frequency distortion is made negligible by the use of special devices at the repeater stations for correcting the attenuation-frequency distortion effects.

The velocity distortion is particularly noticeable during the building-up and dying-down periods, when it manifests itself as transient distortion. The duration of transient distortion depends, among other factors, upon the length of the line, the nominal velocity, and the cut-off frequency of the loading. In the old standard loading systems the high frequency velocity distortion caused by the lumpiness effects of the loading was more serious than the low frequency velocity distortion. Accordingly, a substantial reduction in the transient distortion has been obtained in the new standard loading systems by raising the cut-off frequency of the loading.

For further discussion of velocity distortion reference should be made to Mr. A. B. Clark's paper,[19] previously mentioned, which gives experimental results and to an earlier Institute paper by Mr. J. R. Carson [20] which gives the results of theoretical studies.

[18] "Echo Suppressors for Long Telephone Circuits," A. B. Clark and R. C. Mathes, *Jour. A. I. E. E.*, p. 618, June, 1925.

[19] Clark, Loc. Cit.

[20] "Theory of the Transient Oscillations of Electrical Networks and Transmission Systems," J. R. Carson, Trans. A. I. E. E., Vol. 38, 1919, p. 345.

Characteristics of Improved Cable Loading Systems. The principal electrical features of the H-44-25 and H-174-63 phantom group loading systems which have been developed primarily for use on long repeatered cables are given in Table X. Corresponding details of the older standard loading system developed for non-repeatered cables are also included in this table. Typical attenuation-frequency curves of the old and new loading systems are given in Fig. 10.

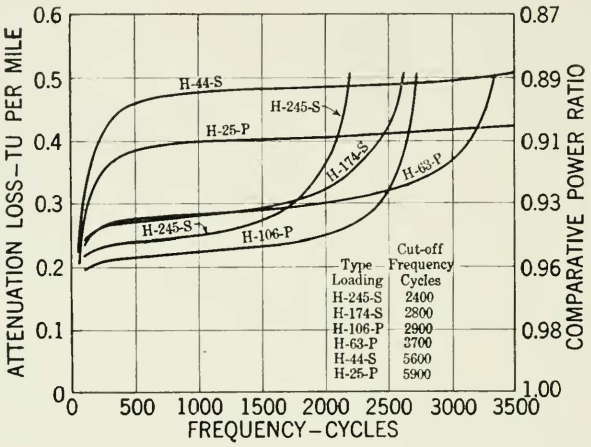


Fig. 10—Attenuation-frequency characteristics of toll cable loading

In the following discussion of detailed characteristics, the various phantom group loading systems will be referred to in terms of recently standardized designations which include a letter to symbolize the coil spacing, in combination with two numbers which correspond with the effective coil inductance values in milhenrys, the first number referring to the side circuit coils and the second number to the phantom coils. The individual side circuit or phantom circuit loading systems have designations which include a letter to symbolize the coil spacing, coupled with the inductance value of the loading coils in milhenrys, and having a letter suffix "*S*" or "*P*" indicating the type of circuit, side or phantom.

The fundamental differences between the new and the old loading systems are with respect to velocity of wave propagation and cut-off frequency, these changes having been made in accordance with the preceding discussion primarily for the purpose of reducing echo effects and transient distortion. For reasons of plant economy and

TABLE X

Loading Systems—Small Gage Repeatered Toll Cables

Item	(*a*) Loading System	Circuit	(*b*) Coil Code No.	Nominal Impedance (Ohms)	Nominal Cut-off Frequency (Cycles)	Transmission Velocity Miles per Second	(*c*) Attenuation Loss TU per Mile at 1000 Cycles		(*d*) Maximum Geographical Length (Miles)
							19 A. w. g.	16 A. w. g.	
(1)	H- 44-25	Side	590	800	5600	19000	0.48	0.25	5000
(2)	"	Phantom	591	450	5900	20000	0.40	0.21	5000
(3)	H-174-63	Side	584	1550	2800	10000	0.28	0.16	500
(4)	"	Phantom	587	750	3700	13000	0.28	0.16	1500
(5)	H-174-106	Side	584	1550	2800	10000	0.28	0.16	500
(6)	"	Phantom	583	950	2900	10000	0.22	0.13	500
(7)	H-245-155	Side	582	1850	2400	8000	0.25	0.16	250
(8)	"	Phantom	581	1150	2400	8000	0.20	0.12	250

NOTES: (*a*) Nominal coil spacing is 6000 feet in cable having a capacitance of 0.062 μf/mile in the side circuits and 0.100 μf/mile in the phantom circuits.

(*b*) The loading coil data are given in Table IX.

(*c*) These attenuation values apply at 55 deg. fahr. Under extreme temperature conditions, the actual attenuation may be approximately 12 per cent larger or smaller, due principally to changes in conductor resistance with temperature. In long repeatered cable circuits these variations of attenuation with temperature require special corrective treatment by means of automatic transmission regulators. (Reference No. 10.)

(*d*) These length limitations are set by transient distortion effects; echo currents may limit circuit lengths to lower values, depending on the grade of balance of the lines and the permissible over-all loss.

flexibility, the new loading systems all have the same coil spacing of 6,000 feet.

The coil spacing being fixed, it necessarily follows that any reduction in coil inductance for the purpose of raising the cut-off frequency will also increase the transmission velocity. The attenuation improvement obtained by the loading decreases as the velocity is increased. High velocity loading is more expensive than low velocity loading, in the sense that more repeaters are required for the same over-all loss. Obviously, although high velocity loading could be used for short haul traffic, it would not be so economical as a low velocity loading. Commercial considerations thus justify a series of loading standards, graded to meet the requirements of the different lengths of circuits.

At the present time the two phantom group loading standards, H-44-25 and H-174-63, are sufficient to meet the graded requirements of commercial toll cable circuits, when used with suitable combinations of conductor sizes and repeaters. Three different general types of repeaters are used, known as the 21, 22, and 44 types.[21] The 21 type is used on two-wire circuits requiring only one repeater, under conditions where switched connections involving other repeaters are not involved. The 22 type is used on two-wire circuits requiring one or more repeaters. The 44 type is used on four-wire circuits, where one pair of wires is used for one-way transmission in one direction and the other pair of wires for transmission in the opposite direction. When phantom circuits are worked on a four-wire basis, each one-way transmission path actually uses four wires.

Table XI lists the combinations of loading, conductor gage, and type of repeater circuit which are used in meeting the wide range of commercial requirements. The position of the facility item in the table indicates the sequence of transmission excellence, Item (i) being the highest grade facility in this respect. In general, the cost of these facilities is in reverse order to the sequence of electrical excellence.

The exact limits of the field of use of a given type of facility depend upon the magnitude of the permissible over-all transmission loss, and upon the grade of repeater balance obtainable. A discussion of these features would bring in complicated engineering questions beyond the scope of the present paper. So far as loading features are concerned, it is sufficient to state that H-44-25 loading is generally used on circuits of approximately 500 miles or more. On circuits intended for switched business, it is frequently necessary to use this

[21] Gherardi—Jewett, Loc. cit.

TABLE XI

Types of Toll Cable Facilities

Item No.	Length Circuit	Cable Gage	Type of Loading	Type Circuit	Type Repeater
(a)	(short)	19	H-174-63	2-wire	—
(b)		16	"	"	—
(c)		19	"	"	21
(d)		16	"	"	21
(e)		19	"	"	22
(f)		16	"	"	22
(g)		19	"	4-wire	44
(h)		16	H- 44-25	2-wire	22
(i)	(very long)	19	"	4-wire	44

type of loading for much shorter distances. For further discussion of the use of repeatered loaded lines reference is made to recent papers presented before the Institute by Mr. J. J. Pilliod [22] and Mr. H. S. Osborne.[23]

H-63-P versus H-106-P Loading. The standardization of the H-63-P loading to replace the H-106-P loading for association with H-174-S loading, is of particular interest in illustrating the reactions of repeater requirements on loading design. Phantom circuits necessarily have a lower attenuation constant than the associated side circuits, when the loading is designed to meet the same standard of cut-off frequency and the coils are spaced at the same loading points. When repeaters are used on such loaded phantom circuits, the net equivalent is practically no lower than the net equivalent of the associated side circuits, due principally to the fact that the loaded sides and phantoms have practically the same velocity and cut-off frequency characteristics.

Under present operating conditions for short small gage loaded circuits of such lengths that satisfactory transmission results can be obtained without using telephone repeaters, there is ordinarily no important advantage in having the phantom circuit more efficient than the side circuits. It is a distinct operating convenience, of course, to be able to use the phantom circuit and its associated side circuits indiscriminately for the same class of service.

Having the above situations in mind, it was decided to redesign the phantom loading so that it would have approximately the same

[22] "Philadelphia-Pittsburgh Section of New York-Chicago Cable," J. J. Pilliod, Trans. A. I. E. E., Vol. 41, 1922, p. 446; *Bell System Technical Journal*, Jan., 1922.

[23] "Telephone Transmission over Long Distances," H. S. Osborne, Trans. A. I. E. E. Vol. 42, 1923, p. 984.

attenuation constant at 1,000 cycles as the associated H-174-S loading. This resulted in the reduction of the phantom loading coil inductance to 63 milhenrys. On the basis of equal attenuation losses in the phantom circuit and its side circuits, the continued use of a higher grade coil in the phantom circuit was no longer justified from a cost standpoint. Accordingly, the new 63-milhenry phantom coil (Code No. 587, Table IX) was designed to have approximately the same d-c. resistance as the earlier standard 106-milhenry coil (Code No. 583), since this permitted a substantial reduction in the size of the loading coil and a consequent reduction in cost, without increasing the over-all losses in the associated side circuits. The design finally chosen resulted in the phantom coil having approximately the same over-all dimensions as the associated side circuit coils. This permitted the phantom coils to be mounted on the same spindles with the associated side circuit loading coils as phantom groups, thus reducing the amount of inside cabling. This gave improved electrical results, besides reducing the potting costs. The use of the smaller size phantom coil, in combination with a larger size case, made it practicable to pot a total of 45 phantom group combinations (135 coils) in a single case. Using the same size case for potting phantom group combinations involving the older large size phantom coils, the limit on the number of coils was 108 (36 phantom groups).

The reduction of the phantom coil inductance from 106 to 63 milhenrys made a substantial increase in the cut-off frequency and in the velocity of transmission, as noted in Table X. These improved characteristics made the H-63-P circuit much superior to the H-106-P circuit from the standpoint of echoes and velocity distortion characteristics. On this basis the H-63-P circuit is intermediate in transmission excellence between H-174-S and H-44-25 circuits.

It was found inadvisable to make a similar change in the H-44-25 loading system owing to cross-talk reactions following from the necessary use of higher repeater gains in the phantom circuit. These undesirable reactions, though present to a lesser degree in the case of the H-174-63 system were offset by the factors already described. The size of the H-25-P coil was, however, reduced to comform to the potting method adopted for H-174-63 loading.

From the standpoint of repeater circuits the H-174-63 system is inherently better than the H-245-155 system because of its higher velocity and higher cut-off, with resulting higher quality of transmission. Furthermore, as far as non-repeatered circuits are concerned, there is a negligibly small difference between the transmission performances, considering frequency distortion effects as well as volume

efficiency effects. The standardization of the H-174-63 phantom group loading system, therefore, marked the abandonment of use in new facilities of the old standard H-245-155 phantom-group loading system.

Attenuation—Frequency Distortion. In addition to their improved velocity and cut-off frequency characteristics, the H-44-25 and H-174-63 loading systems have an important advantage from the standpoint of attenuation-frequency distortion effects, as is illustrated in Figs. 10 and 11. The frequency distortion effects illustrated in Fig. 10

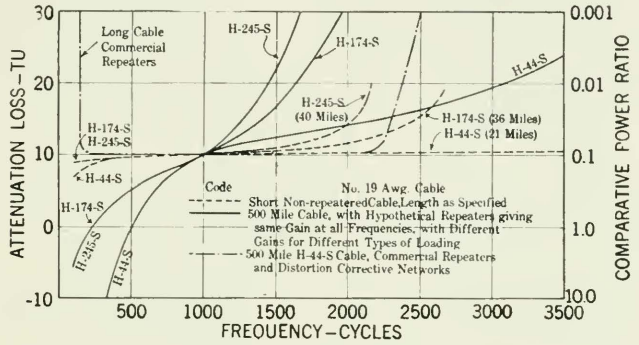


Fig. 11—Attenuation-frequency characteristics of short and long loaded toll cable circuits having a net attenuation loss of 10 TU at 1000 cycles

may become very serious in very long lines. An indication of this is given in Fig. 11. The heavy line curves in this diagram illustrate the attenuation-frequency characteristics of a 500-mile 19 A.w.g. cable circuit involving the various types of loading noted, assuming that "perfect repeaters" are used in each case to reduce the total line loss to 10 TU at 1,000 cycles. The foregoing "perfect repeater" is assumed to have the same amplification at all frequencies. Of course, in order to have the same over-all efficiency in the different types of circuits at 1,000 cycles, it is necessary to assume different total amounts of repeater gain. The dotted lines in Fig. 11 illustrate corresponding frequency characteristics of short non-repeatered cables having the same types of loading as before; in each case the length of 19 A.w.g. cable circuit being chosen so that the non-repeatered circuits have the same loss (10 TU) at 1,000 cycles. A visual inspection of the dotted and heavy line curves indicates how the line losses pile up in long connections. In the old standard low cut-off loading, the accumulated losses in very long lines amount to a suppression effect for frequencies above 1,600 cycles.

In very long lines having the newer grades of loading, the line losses are still sufficient to cause serious attenuation distortion effects if allowed to go uncorrected. The improved types of repeaters now used on long loaded circuits provide somewhat higher gains at the upper speech frequencies, thereby obtaining approximately a flat frequency characteristic over a wider frequency range. In repeaters used in conjunction with the H-44-25 loading, losses are introduced at the lower speech frequencies by auxiliaries to the repeater circuit, for the purpose of flattening the frequency characteristic at low frequencies. An indication of the improvement obtainable in the above ways is given by a dot-dash curve in Fig. 11, which illustrates the attenuation-frequency characteristic of a 500-mile H-44-S circuit having the best types of repeaters now commercially available.

In view of the difficulties brought into repeatered circuits by the use of loading, the question comes up: "Why not use more repeaters and do without the loading?" In the case of long cable circuits the answer to this question is that the coil loading substantially improves the attenuation and substantially reduces the frequency distortion at a cost which is much lower than the cost of the additional repeaters and distortion corrective networks which would be required to give the same grade of transmission without using loading.

Long Repeatered Open Wire Lines. In the case of the long open wire lines, the present day answer to the foregoing question is unfavorable to the use of loading. The use of improved types of repeaters now makes it possible to secure better transmission results in long repeatered circuits without loading, than can be secured in loaded repeatered lines. In this connection it should be noted that in the case of non-loaded open wire lines the distributed inductance is sufficiently large to keep the attenuation-frequency distortion low. Also the velocity of transmission is very high relative to that of a coil loaded line and there is no cut-off effect except that produced by the filters and other apparatus in the repeater sets.

These general transmission considerations are resulting in the removal of coil loading from high grade open wire lines. This dismantling work is being accelerated in order to adapt the open wire plant for a much more extensive application of carrier telephone and carrier telegraph systems.

The present expectations are that in the future new applications of open wire loading will generally be limited to isolated cases of short lines where carrier telephone or telegraph systems are not contemplated and where the maintenance and operating conditions are unfavorable to the use of telephone repeaters.

Cable Loading Installation Features. Cost considerations make it desirable to use aerial cable in the long toll cable installations, so this type of construction is generally used in the open country. In the vicinity of large population centers, underground cable is used.

Typical aerial cable loading installations are illustrated in Figs. 12 and 13. On the main trunk cables two-pole *H* fixtures capable of supporting four to six large coil cases are usually required. Fig. 12

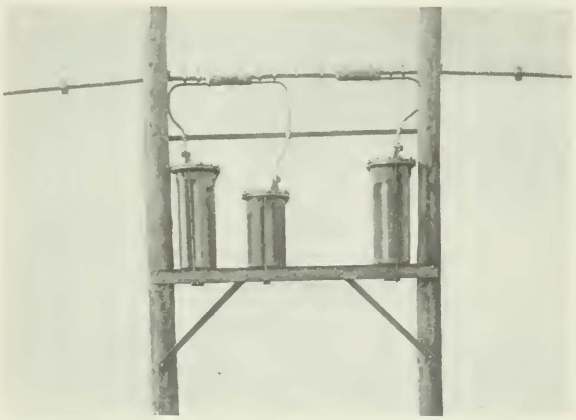


Fig. 12—Installation of aerial toll cable loading on 4-case "H" fixture

illustrates a fixture of this type designed for supporting four cases, three of which are already in place. On the smaller branch cables a single pole fixture such as illustrated in Fig. 13 is commonly used.

At the time a toll cable is installed, provision is made in the cable splices for the ultimate requirements as well as for the initial loading installation. Ordinary splices are made for the coils which are installed at the time the cable is placed, and "balloon" splices which provide the slack wire required for splicing are arranged for subsequent installations.

III. Loading for Incidental Cables in Open Wire Lines

In the loading applications discussed in the preceding sections, the primary purpose of the loading is to reduce line attenuation losses and frequency distortion effects. In the case of incidental pieces of cable in open wire lines, however, the primary function of the loading is to give the inserted cable approximately the same impedance characteristics as the open wire line, in order to minimize reflection effects

at the junction of the cable and the open wire construction. An incidental cable occurring at a line terminal is ordinarily known as a toll entrance or a terminal cable; when occurring at an intermediate point, it is known as an intermediate cable.

The reduction of junction impedance irregularities has become especially important during recent years as a result of the rapidly increasing use of telephone repeaters, since in repeatered circuits,

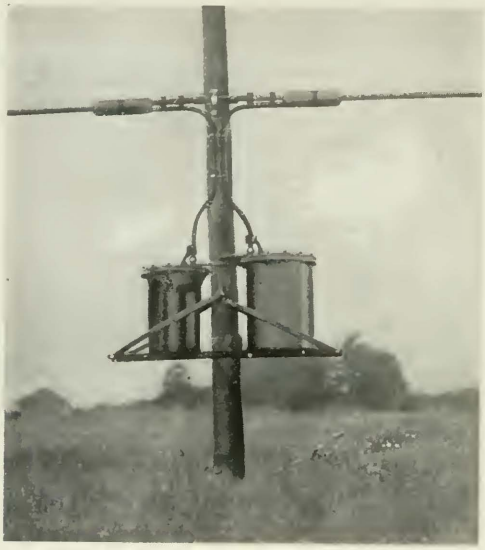


Fig. 13—Installation of aerial toll cable loading—single pole fixture for small branch cables

line impedance irregularities, by virtue of their effect upon the repeater circuit balance, may reduce the effective repeater gain and thereby impair transmission by an amount much larger than the ordinary reflection loss. Prior to the general use of telephone repeaters, satisfactory results were obtained by using some one of the standard heavy or medium weight cable loading systems on the entrance and intermediate cables associated with loaded open wire lines, and a special weight of loading was used on the incidental cables in the non-loaded open wire lines. In some cases ordinary medium loading was used, with suitable types of step-up or step-down transformers at the terminals of the inserted cable.

Incidental Cables in Loaded Open Wire Lines. In toll entrance and intermediate cables associated with loaded open wire lines, the

primary requirements for matching impedance are that the nominal impedance and the cut-off frequency of the cable loading and of the loaded open wire line should be closely the same. To a first degree of approximation this means that the cable loading sections should have the same total mutual capacitance as the open wire loading sections, which, of course, requires a very much closer spacing. The cable loading system which was standardized for use in association with loaded open wire lines is designated "E-248-154". Its primary electrical characteristics are given in Table XII. Besides meeting

TABLE XII

Typical Loading Systems for Toll Entrance and Intermediate Cables

Loading System Designation	Type Circuit	Coil Inductance Henrys	Coil Spacing Miles	Nominal Impedance Ohms	Cut-off Frequency Cycles	Attenuation Loss TU per Mile at 1000 Cycles
E-28-16	Side	0.028	1.09	650	7200	0.15 } (13 A.w.g.)
	Phantom	0.016	1.09	400	7800	0.13 }
CE-4.1-12.8	Side	0.0041	0.176	600	45000	0.22 } (13 A.w.g.)
	Phantom	0.0128	1.09	400	8500	0.19 }
M-44-25	Side	0.044	1.66	650	4600	0.29 } (16 A.w.g.)
	Phantom	0.025	1.66	400	4900	0.24 }
E-248-154	Side	0.248	1.09	1950	2400	0.081 } (13 A.w.g.)
	Phantom	0.154	1.09	1200	2500	0.070 }

NOTE. Cable capacitance is assumed to be 0.062 μf per mile for side circuits, and 0.100 μf per mile for phantoms.

the impedance requirements for use in association with repeatered open wire lines, it is also very satisfactory with respect to attenuation characteristics. In placing this loading, it is customary to locate the first loading point in the cable at such a distance from the last loading point in the open wire line that the total capacitance of the junction loading section is closely the same as that in the regular open wire loading sections.

Incidental Cables in Non-Loaded Open Wire Lines. The problem of designing coil loading for incidental cables in non-loaded open wire lines is considerably more complicated than the case above discussed, primarily because it involves an impedance match between a smooth line and a lumpy line. Broadly stated, the first part of the problem is to design a loaded cable of such characteristics that its

corresponding smooth line is closely similar to the non-loaded open wire line. The second and more complicated part of the problem is to determine the coil spacing. This usually involves some degree of compromise, because of the dependence of the impedance of a loaded cable upon the loading termination.

The first general requirement is that the ratio of inductance to capacitance to resistance per unit length in the loaded cable should be the same as the corresponding ratio in the non-loaded open wire line. Ordinarily, the loading coil resistance does not play an important part in the determination of the optimum resistance for the loaded cable, the choice of conductor gage being far more important. From this point of view, No. 13 A.w.g. is practically the best gage of conductor for entrance cable circuits connecting with 165-mil open wire lines. For the optimum impedance match on cables connecting with 104-mil open wire lines, it is necessary to use much higher resistance conductors, the choice between Nos. 16 and 19 A.w.g. conductors depending upon a number of factors which space limits do not allow to be discussed.

As noted in the discussion under "Theory" in the first part of the paper, the characteristic impedance of a uniform line is substantially a pure resistance, having the value $\sqrt{L/C}$ over the frequency range throughout which the inductive reactance per unit length is large with reference to the resistance. On the other hand, the characteristic impedance of a coil loaded cable varies over a wide range with frequency, depending upon the particular loading termination used.

Typical impedance-frequency curves for mid-coil and mid-section terminations are illustrated in Fig. 3. As will be seen from this diagram, the rising slope of the mid-section termination and the drooping slope of the mid-coil termination do not deviate greatly from a straight line relation for frequencies below approximately 0.5 of the cut-off frequency. The higher the cut-off frequency is, the more closely will the impedance-frequency characteristic of the loaded cable approach the flat characteristic of the non-loaded open wire line over the range of frequencies involved in speech transmission. In this connection, it is to be noted that the repeaters now used on open wire lines are designed to transmit frequencies between approximately 200 and 2,600 cycles. Of course, the higher the cut-off frequency, the more expensive will be the loading. Practical reasons make it desirable to space the loading coils on the cable circuits connecting with non-loaded open wire lines at the same intervals as the coils which are used on the cable circuits connected with the loaded open wire lines. This consideration in combination with the nominal impedance

requirement previously mentioned, fixes the cut-off characteristics and, hence, the slope of the termination impedance-frequency characteristic.

These general considerations have led to the standardization of the E-28-16 loading system for use on entrance cable and intermediate cable conductors associated with non-loaded open wire lines. General data for this system are given in Table XII, and the computed impedance characteristics are illustrated in Fig. 14, which also gives

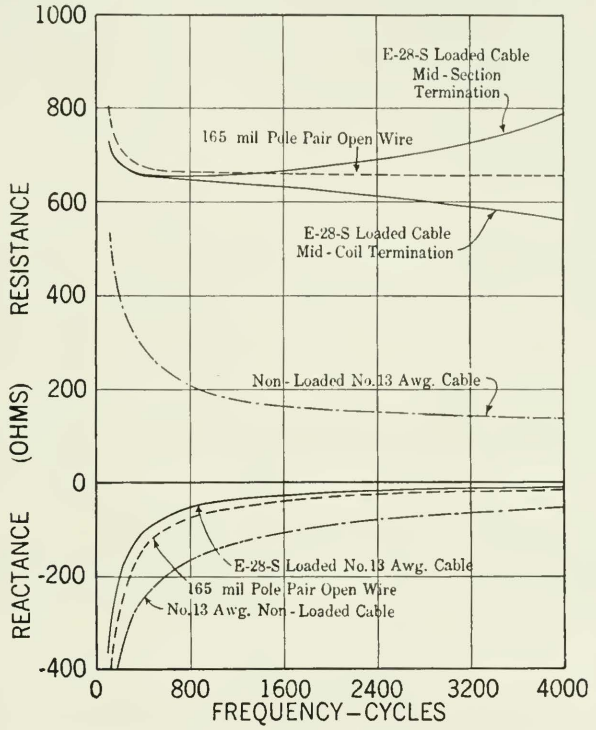


Fig. 14—Typical impedance-frequency characteristics of loaded and non-loaded entrance cable and non-loaded open wire line

the characteristic impedance curves for the non-loaded open wire line and the non-loaded cable. Since the E-28-16 loading system is a low impedance loading, the attenuation improvement is small relative to that of other types of loading system which are primarily installed for attenuation improvement.

Table XII also gives general data regarding the M-44-25 entrance cable loading system which has been used to some extent as a substitute for the E-28-16 loading system on cables connected to non-loaded open wire lines. The M-44-25 system used higher inductance

loading coils and considerably longer spacing intervals than the E-28-16 system, and was consequently less expensive. The impedance characteristics, however, were not so satisfactory at the upper speech frequencies because of the greater slope of the impedance-frequency characteristic, due to the lower cut-off.

Carrier Frequency Loading. Special types of entrance cable loading have been developed for use on incidental cables in open wire lines on which carrier telephone or carrier telegraph systems are superposed. Loading system CE-4.1-12.8 listed in Table XII exemplifies this type of loading. The present standard carrier telephone systems operate up to frequencies of the order of 30,000 cycles.[24]

In order to get satisfactory impedance and attenuation characteristics in the loaded incidental cables, a cut-off frequency of approximately 45,000 cycles is used.

The highest working frequency in carrier loaded cables is approximately 0.75 of the cut-off frequency. The ordinary mid-coil and mid-section terminations do not give sufficiently close approximations to a flat impedance-frequency characteristic over this wide range of frequencies, so it has been necessary to use at the terminals of carrier loaded cables, a simple impedance corrective network.

Data regarding attenuation losses in a typical carrier loaded cable are given in Table XIII. For purposes of comparison similar data are given on a corresponding non-loaded cable. Effective resistance values of the carrier loading coil are also included.

The high frequency loading is used only on the side circuits, since at the present time it is not customary to operate carrier telephone systems over phantom circuits. The associated phantom circuit

TABLE XIII

Carrier Frequency Loading

Frequency Kilocycles	Attenuation Loss-TU per Mile (13 A. w. g. Cable)		Resistance-Ohms per Carrier Loading Coil
	Non-Loaded	C-4.1 Loading	
1	0.49	0.23	1.5
5	0.78	0.27	1.6
10	0.90	0.33	1.9
20	1.14	0.52	4.1
30	1.37	0.90	8.1

[24] "Carrier Current Telephony and Telegraphy," E. H. Colpitts and O. B. Blackwell, Trans. A. I. E. E., Vol. 40, 1921, p. 205.

loading is designed for ordinary speech transmission. In order to transmit the high frequency carrier currents over the side circuits, it is necessary to have the side circuit loading coils spaced much more closely than for the ordinary voice frequency loading coils in the phantom circuit. On this account the theoretically best loading points for the carrier circuits frequently occur at places where it is inconvenient to locate the loading coils. The actual loading sections in such cases are made shorter than the theoretical lengths, and the deficiencies in loading section capacitance are remedied by adding lumped capacitances in the form of "building-out condensers." Recently, special types of stub cable designed specially for building out purposes have come into use as substitutes for building-out condensers.

Loading Coils. The design of the coils used in the E-28-16 and M-44-25 loading systems is generally similar to the toll cable loading coils having 35-permeability compressed powdered iron cores already described. The loading coils used in the E-248-154 loading system are larger coils of the air-gap type 65-permeability wire core construction listed in Table VII.

As regards the carrier loading system, CE-4.1-12.8, since this involves the transmission through the loading coils of frequencies up to 30,000 cycles or somewhat higher, special coil designs are required. The coil which loads the audio frequency phantom circuit, aside from being specially balanced for association with the side circuit coils, is generally similar in construction to the compressed powdered iron core phantom coil for toll cables.

The side circuit coil, however, is used for loading the high frequency circuit, and more severe requirements are, therefore, imposed on it owing to the multi-frequency transmission. Ordinarily the circuits are equipped to provide three or four carrier telephone channels or 10 carrier telegraph channels over a pair of wires, in addition to the ordinary audio frequency telephone and grounded telegraph channels. The primary added requirements as regards the loading coils are freedom from intermodulation between channels, and low energy losses at carrier frequencies. The most satisfactory solution as regards freedom from magnetic modulation is the avoidance of the use of ferro-magnetic core materials. The side circuit loading coils were, accordingly, designed as toroidal wood core coils, with finely stranded copper windings in order to limit the eddy-current losses. Data regarding resistance-frequency characteristics are included in Table XIII.

The air core side circuit coils have a small leakage inductance which must be allowed for in determining the phantom coil inductance. For this reason the phantom coil inductance is lower than in the E-28-16 system (Table XII.) In order to avoid impedance irregularity in the carrier circuits at the phantom loading points, it is necessary that the combination carrier-phantom loading units should have closely the same total inductance and shunt capacitance as the ordinary carrier loading coils. This requires the use of a different type of carrier loading coil at the phantom loading point from that at the non-phantom loading points, having a lower inductance and capacitance corresponding to the leakage inductance and shunt capacitance of the associated phantom coil. Other refinements of design are involved in these combination loading units.[25]

IV. Cross-Talk

One of the greatest practical difficulties which has been encountered in extending the commercial range of long distance telephone service is that of keeping at a tolerably low value, the speech overhearing effects known as cross-talk, which occur between adjacent telephone transmission circuits whenever there is an appreciable amount of electromagnetic or electrostatic coupling between them.

From the early days of telephony great care has been exercised in plant design and construction work to avoid circuit and apparatus unbalances, but as is to be expected from the nature of the problem, it is practically impossible to obtain and maintain absolutely perfect balance. In short telephone circuits, there is no particular difficulty in keeping the unbalance effects small enough so that the over-all cross-talk is not serious. As the length of the line increases, however, there are more and more opportunities for unbalances in the lines and in the associated apparatus in the lines and offices. In repeatered lines, moreover, the repeaters amplify the cross-talk as well as the speech transmission. Thus we have the cumulative effects of cross-talk from successive sections in the long repeatered lines. From the service standpoint, moreover, it is necessary that the cross-talk in the very long lines should be within the limits set for the shorter lines.

The problem of keeping cross-talk low between a phantom circuit and its associated side circuits, and between the two associated side circuits of a phantom group, is by far the most difficult phase of the general cross-talk problem in long repeatered cables. It is present in the cables, the loading coils, the terminating apparatus and the office

[25] U. S. Patents Nos. 1,501,959, Martin and Shaw; 1,501,926, Shaw.

cabling. Of these, the cable and associated loading coils are the major sources of unbalances.

The phantom-to-side and side-to-side cross-talk unbalances in the cable quads are reduced to small values by exercisng great care both in the various manufacturing processes and in the selection of raw materials. When the cable is installed in the field, a large improvement in cross-talk conditions is secured by splicing adjacent lengths of cable together in such a way that the unbalances in one length of cable substantially neutralize the unbalances contributed by the adjacent length of cable. Usually, three such "capacity-unbalance test" splices are made at symmetrical points in each loading section and as a result the average over-all capacity unbalance in a loading section is reduced to about one-tenth of the magnitude which would hold if these test splices were not made.

In the design of the standard phantom circuit and side circuit loading coils, special care was taken to make them substantially free from inherent unbalances. Also in the manufacture of the coils, great care is exercised to realize the benefits of the inherent symmetry of the designs. In the early days before telephone repeaters came into general use on loaded lines, satisfactory results from the standpoint of self inductance and mutual inductance unbalances were obtained by adjusting the different windings to the nearest turn; *i.e.*, a condition of balance where either adding or subtracting one turn to one of the line windings would increase the cross-talk rather than reduce it.

Later when repeaters came into general use, it was found necessary to obtain much more refined adjustments. Further improvements have been worked out in manufacturing methods and processes which allow a greater degree of symmetry. As a result of these various improvements, the phantom-to-side cross-talk unbalances in the loading coils have been reduced approximately 75 per cent. or more below the values obtained before repeaters came into general use on small gage toll cable. The coil cross-talk unbalances are now nearly as low as the cross-talk unbalances in the associated cable sections after the completion of the capacity unbalance test splicing.

The loading coils used in the very long circuits having H-44-25 loading obviously are more important from the standpoint of cross-talk limitations than the coils used in the shorter circuits having H-174-63 loading, and somewhat greater care is required in their manufacture. These coils are adjusted and tested in a factory test circuit which at the cross-talk test frequency simulates the service impedance conditions. In the phantom-to-side cross-talk test, the disturbing test current is superposed on the phantom circuit, and

measurements of the cross-talk are made in the side circuits, the cross-talk being expressed in millionths of the current into a transformer connected to the phantom circuit and of such ratio as to make the impedance at its input equal to that of the side circuit. As a result of the improvements previously mentioned, the average cross-talk in the coils used for the H-44-25 loading is now about 20 millionths. This corresponds to an attenuation of about 95 TU.

To assist in visualizing the real achievement which this minute value of phantom-to-side cross-talk represents, Table XIV gives information regarding the cross-talk of different elementary types of unbalance in H-44-25 loading coils:

TABLE XIV

Cross-talk Due to Unbalance in H-44-25 Loading Coils

Type of Unbalance	Amount of Cross-talk
1 ohm resistance	400 millionths (68 TU)
1 micro-henry inductance	2.5 " (112 TU)
1 turn of winding	280 " (71 TU)
1 micro-microfarad capacitance	0.94 " (121 TU)

These values apply at 1,000 cycles.

In the loading coils designed for H-174-63 loading, the cross-talk per unit of electromagnetic unbalance tends to be smaller and the cross-talk per unit of electrostatic unbalance larger, in rough proportion to the differences in line impedance between the H-44-25 and H-174-63 circuits.

Side-to-side cross-talk is uniformly lower than phantom-to-side cross-talk, as would be expected from the less intimate coupling between circuits. Accordingly, the special adjustments which are made are primarily for the purpose of reducing phantom-to-side cross-talk.

In the loading coils intended for H-44-25 circuits the special cross-talk adjustments are applied for minimizing "far-end" cross-talk or for minimizing "near-end" cross-talk, according as the coils are required for four-wire or two-wire repeatered circuits, respectively. The term "far-end" cross-talk applies to cross-talk heard at the distant end of the disturbed circuit, and correspondingly the term "near-end" cross-talk applies to the cross-talk heard at the end of the disturbed circuit near the talker.

Considering now the cross-talk between four-wire circuits in the same quad, it is to be noted that the directional effects of the telephone repeaters block the transmission of cross-talk in the one-way path back to the near end of the circuit, and consequently the special cross-talk adjustments on the coils for four-wire H-44-25 circuits are made primarily for reducing far-end cross-talk.

In two-wire circuits, near-end and far-end cross-talk both occur, and generally near-end cross-talk is much greater because its "average" cross-talk path has less attenuation than that of the far-end cross-talk. Consequently, the special cross-talk adjustments made in the two-wire circuit coils are for the purpose of reducing the near-end cross-talk to a minimum.

In the foregoing connection, it is to be noted that the cross-talk current caused by electromagnetic unbalances flows around the two ends of the disturbed circuit in series. On the other hand, the cross-talk current caused by electrostatic unbalances divides and flows from its point of origin in opposite directions around the two ends of the circuit in parallel. Consequently, when electrostatic and electromagnetic cross-talk currents are in phase at one end of the circuit, they will be practically in phase opposition at the other end of the circuit. The special cross-talk adjustments are made in such a way as to get the maximum benefit from the phase opposition at the particular end of the circuit where the reduction is more important.

In the four-wire type of circuit used in very long cable circuits, relatively large amplification gains are possible in the repeaters because of the characteristic circuit feature which allows the repeaters to act as one-way amplifiers. As a result of these high amplifications, there are large differences in power level on the input and output sides of the repeaters. This fact has made it desirable for cross-talk reasons to segregate the oppositely transmitting branches of the four-wire circuits. In the cables, the "east-bound" and "west-bound" branches of the four-wire circuits are in different groups. This segregation is also carried out in the loading coil pots, and in the office cabling.[26]

With loading coils as manufactured at present, the cross-talk unbalances in the loaded cables are such that the resultant over-all cross-talk is expected to be tolerable for the longest circuits now definitely planned in cable. The margin below commercial limits is much less in two-wire circuits than in four-wire circuits. At present, there is a growing tendency to use two-wire circuits for longer distances

[26] U. S. Patent No. 1,394,062—O. B. Blackwell.

than formerly, for reasons of plant economy. This trend thus increases the severity of the cross-talk requirements.

Unbalances in loaded circuits which contribute to noise due to induction from power transmission and distribution circuits are similar in nature to those contributing to cross-talk. The precautions which are taken in the design, manufacture, and installation of loaded circuits to reduce unbalances have the effect, therefore, of reducing both cross-talk and noise.

V. Telegraphy Over Loaded Telephone Circuits

It had been the practise in the Bell System, before the advent of loading, to employ circuits for simultaneously transmitting telephone and telegraph currents. Two methods were in general use, (1) the composite system, in which each line wire of the telephone circuit provided a telegraph channel with ground return, and (2) the simplex system, in which the two conductors in parallel were used with a ground return.

It was very desirable to continue to superpose d-c. telegraph currents on telephone circuits after the introduction of loading. The possible detrimental effects of the superposed telegraph and telephone currents passing through the loading coils did not require serious consideration so long as the circuits were relatively short, since the magnetic modulation in the loading coil cores due to superposed hysteresis effects was sufficiently small to be negligible. As the length of the loaded circuit was increased, the interaction between the telegraph and telephone currents which has been designated in an Institute paper [27] as "flutter," was aggravated and serious distortion of speech resulted.

Measurements of flutter effects obtained with the two grades of core material then in use, viz., 65-permeability and 95-permeability wire, showed the lower permeability core material to be substantially better in this respect. This material has already been adopted for the high efficiency loading coils used on the large gage toll cable circuits, and for open wire lines used in spanning considerable distances.

In order to obtain improved transmission over composited and loaded Nos. 16 and 19 A.w.g. cables, the side circuit and phantom loading coils for this grade of service were redesigned in 1913 to employ 65-permeability cores working to the same over-all dimensions. In

[27] "Hysteresis Effects with Varying Superposed Magnetizing Forces," W. Fondiller and W. H. Martin, Trans. A. I. E. E., Vol. 40, 1921, p. 443.

addition to improved flutter characteristics, the replacing loading coils had somewhat lower iron losses, and their cost was slightly higher.

A substantial reduction in flutter distortion effects was later obtained in superposed telegraphy over loaded open wire lines and long coarse gage cable circuits, with the adoption of the air-gap type of loading coils already described (Table VIII) as the latter had considerably better hysteresis characteristics than the corresponding types of continuous wire core coils which they superseded.

The operating requirements for the grounded telegraph systems referred to above, necessitated the use of telegraph currents of very large amplitude relative to the telephone currents; consequently on such circuits it was impracticable to realize the benefits of reduced flutter distortion which would have resulted from the use of small amplitude telegraph currents. These possibilities, however, have been fully realized by the development of a metallic polar duplex telegraph system [28] to meet the special requirements imposed by superposed telegraph operation over long small gage telephone circuits. In this system, the superposed telegraph current is of the same order of magnitude as the telephone current. Under these favorable operating conditions, the flutter distortion effects caused by modulation in the cores of the present standard 35-permeability compressed iron powder core loading coils, are within satisfactory limits on the longest circuits which are used simultaneously for telegraph and telephone service.

The recent development of a voice frequency carrier telegraph system [29] providing 10 or more independent channels over a loaded four-wire cable circuit has made it economical to concentrate a large part of the telegraph service over the long repeatered cables on a special group of wires which are not used simultaneously for telephone purposes. This method of operation obviously eliminates all possibility of modulation effects between the carrier telegraph circuits and the speech transmission circuits. However, the possibility of intermodulation effects between the different superposed carrier telegraph channels involves the same fundamental requirements in the loading coils as when the telegraph circuits are superposed on telephone circuits. The requirements of these systems are satisfactorily met by the 35-permeability compressed iron powder core loading coils now standard for use in toll cable loading.

[28] "Metallic Polar-Duplex Telegraph System for Cables," Messrs. Bell, Shanck and Branson, Trans. A. I. E. E., 44:337, 1925.

[29] "Voice-Frequency Carrier Telegraph Systems for Cables," Messrs. Hamilton, Nyquist, Long and Phelps, Trans. A. I. E. E., Vol. 44, 1925, p. 327.

VI. Recent Improvements in Loading for Exchange Area Cables

The developments discussed in the preceding sections were directed to improving and extending the range of long distance telephone service. During the greater part of this period the loading standards for exchange area trunk cables remained fixed. The first important change occurred about 1916, when compressed powdered iron core coils came into general use in place of the old standard wire core coils.

In the period 1922–4, the use of new types of fine wire cables had reached a point which required that certain changes be made in the old standard loading systems. Accordingly a new series of improved loading systems having a considerably higher cutoff frequency than the original standard systems, described in Table II, were developed.

Cable Developments. Notable advances have been made in the art of cable manufacture during the last decade or so, including the standardization of 450-pair 19-A.w.g. cable, 900-pair 22-A.w.g. cable, and 1,200-pair 24-A.w.g. cable, all contained in standard full size sheaths (2⅝ in. outside diameter). For each of the conductor gages involved, each of these new maximum size cables has approximately 50 per cent. more conductors than the previous maximum size cable, typified by the old standard 300-pair No. 19-A.w.g. or 600-pair No. 22-A.w.g. cables.

The newer types of cables have a smaller amount of paper insulation on the individual conductors with a resultant increase in mutual capacitance.

About 1921 the methods of stranding the newer types of fine wire cables (No. 22 and 24 A.w.g.) were changed in order to improve their balance characteristics. These changes made the cables suitable for the application of loading.

The use of the old standard loading systems on the new types of cables would have resulted in an objectionable impairment of quality, due to the reduction of the cut-off frequency resulting from the increased cable capacitance. Also the types of coils available were more expensive than could be justified for permanent standards on the low cost fine wire cables. Accordingly the development of new loading systems and less expensive coils was undertaken.

Determination of New Cut-Off Frequency Standard. The coil design cost-balance study was taken up as one phase of a general transmission-cost study of exchange area transmission, which also included a theoretical investigation of cut-off frequency standards.

In this work use was made of recent investigations of the effect

of variations in the frequency distortion and volume efficiency of a telephone circuit on the capability of the circuit to transmit and reproduce intelligible speech.

In the cost studies, allowance was made for the reduced costs of the new types of cable facilities, and the use of less expensive types of coils proportioned to be in approximate cost-equilibrium with these facilities. On the basis of these new cost relations, it was found that an increase in the cut-off frequency of exchange area loading could be justified. Further studies showed that if the cut-off frequency should be raised materially above 3,000 cycles the increased costs would be large in proportion to the resultant improvement in transmission. On this basis, it was decided to adopt 2,900 cycles as the cut-off frequency standard for the new loading, when used on higher capacitance cables. This corresponds to a cut-off frequency of approximately 3,200 cycles on the older types of low capacitance cables.

New Standard Loading Systems. Having decided upon a new cut-off frequency standard, the next step was to choose coil spacings and inductance values. Obviously, in order to make full use of available loading manholes and vaults, it is desirable to adhere to the established spacing standards. Also, it is desirable to make as much use as practicable of the old standard loading coil inductance values, in order to minimize the expense of rearranging the existing loading to conform to the new standards. Furthermore, there are important advantages in having a graded series of standards. This avoids economic waste otherwise involved in the use of expensive loading on trunks where a less expensive loading is good enough.

The foregoing general considerations resulted in the standardization of certain new loading systems, the principal transmission features of which are summarized in Table XV.

TABLE XV

New Loading Standards for Exchange Area Trunks

Loading Designation	Coil Spacing Feet	Coil Code Nos.	Approx. Cut-off Frequency Cycles	
			High Capacitance Cable	Low Capacitance Cable
M-88	9000	602	2900	3200
H-135	6000	603	2800	3200
H-175	6000	574	——	2800
D-175	4500	574	2900	3200

NOTE. High capacitance cable has approximately 0.083 μf per mile. Low capacitance cable has approximately 0.066 μf per mile.

The M-88 system is especially suitable for the shorter lengths of fine wire trunk cables which constitute the predominating bulk of the exchange area trunk mileage. In longer trunks, the other more expensive loading systems find their field of service. The H-175 system is limited to low capacitance cables because of the lower cut-off effects on high capacitance cables, but has considerable commercial importance because of the large number of low capacitance cables now in the plant.

Table XVI gives general transmission data on typical exchange area trunks using the new loading systems, including also non-loaded trunks. Attenuation-frequency characteristics of some of these trunks are given in Fig. 15. A dotted line curve shows the characteristics

TABLE XVI

Transmission Characteristics of Typical Exchange Area Trunks

Cable Conductor A.w.g.	Capacitance μf./Mile	System	Coil Code No.	Cut-off Frequency Cycles	Circuit Impedance Ohms	Attenuation Loss TU per Mile
24	0.079	Non-loaded	—	—	740	2.2
22	0.083	" "	—	—	570	1.8
24	0.079	M-88	602	2900	900	1.48
22	0.083	M-88	602	2900	990	0.96
22	0.083	H-135	603	2800	1300	0.68
19	0.085	Non-loaded	—	—	400	1.27
22	0.083	D-175	574	2900	1690	0.53
19	0.083	M-88	602	2800	860	0.51
19	0.085	H-135	603	2800	1280	0.38
19	0.066	H-175	574	2800	1640	0.29
19	0.085	D-175	574	2800	1680	0.30
16	0.066	M-88	602	3200	960	0.24
16	0.066	H-135	603	3200	1420	0.20
16	0.066	H-175	574	2800	1640	0.17

NOTE. The impedance and attentuation figures hold at 1000 cycles. Impedance values for loaded circuits assume mid-section termination.

of the old standard medium loading when used on 0.065 μf per mi. No. 19 A.w.g. cable.

The position of the different types of facility in Table XVI indicates in a general way the sequence in regard to costs of the different types of facility; i.e., No. 24-A.w.g. non-loaded cable circuits are the least expensive of those listed and H-175 loaded No. 16-A.w.g. cable the most expensive, the intermediate facilities being correspondingly intermediate in cost.

In general, it will be noted from Table XVI that the facility cost is in reverse order to the transmission efficiency. However, there are exceptions to this general tendency: for instance, non-loaded No. 19-A.w.g. cable is less efficient and more expensive than No. 22-A.w.g.

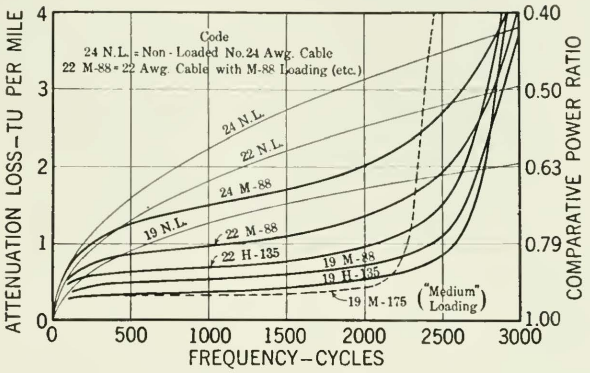


Fig. 15—Attenuation-frequency characteristics of typical exchange area loaded and non-loaded cables

cable with M-88 or H-135 loading. From this it will be apparent that non-loaded No. 19-A.w.g. cable has practically no economical field of service on a competitive basis with loaded No. 22-A.w.g. cable.

The problem of the plant engineer in laying out trunk cables is to determine for each trunk group the type of facility which will meet the established transmission standards with the lowest cost. The permissible loss in a given length of trunk in a particular area depends on the type of service involved; for instance, a larger loss is allowable in a direct interoffice trunk than in other types of trunks forming part of a toll connection.

The allowable loss for a given length and type of trunk varies widely among different metropolitan areas, due to local conditions. These limits are established by means of "loop-and-trunk" studies which determine for a particular area the most economical allocation of the permissible over-all loss between the subscriber loops and the interoffice trunks. On account of the wide range of local conditions the fields of service of the different types of loaded and non-loaded trunks cannot be sharply defined. An indication of the service uses is given in the upper part of Fig. 16, which illustrates the possible applications of the new standard types of facilities for direct interoffice trunk service, assuming maximum allowable losses of 11 and 15 TU, respectively. The diagram shows, for instance, that M-88

loading on No. 22-A.w.g. cable is the preferred construction for trunks from 7.5 to 11.5 miles long, when working to an 11-TU limit. The lower part of Fig. 16 indicates on a cumulative-percentage basis the distribution of direct interoffice trunk lengths in the Bell System.

In the design of the exchange area trunk plant, it is, of course, necessary to consider the signaling characteristics of the facilities and associated equipment, as well as the transmission characteristics.

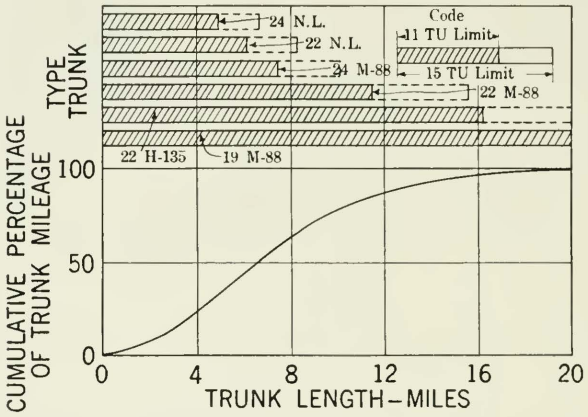


Fig. 16—Direct interoffice trunks

Field of use of different types of loaded and non-loaded facilities, working to 11 to 15 TU limits on attenuation loss. Curve in lower part of diagram shows distribution of trunks with respect to trunk length

After a certain length is reached in a given type of facility, it may become necessary to use relatively expensive signaling equipment for working longer distances. In some cases of this kind, the total facility cost may be reduced by using a more expensive grade of circuit which will allow less expensive signaling equipment.

In the application of the new standard loading systems, the same standards of over-all attenuation loss are adhered to, as in the older loading systems. In consequence, there is an appreciable improvement in the intelligibility of transmission, due to the ability of the new loading systems to transmit efficiently a range of high frequency voice over-tones which are suppressed by the old standard 2,300-cycle cut-off.

Along with this improvement in service, the new loading systems substantially reduce the plant cost; partly due to the economies which result from the extension of the transmission range of the new types of fine wire cables, and partly because of the use of materially less expensive types of loading coils.

Loading Coils and Cases. As previously noted, the first important change in the coils used for exchange area loading from the early 95-permeability wire core type was the substitution of compressed powdered iron in place of wire for the cores. Initially, only coils having powdered iron cores with a permeability of 60 were designed, as this value corresponds to the effective value of the cores displaced. More recently, in order to better fit in with the requirements of the new cut-off frequency standard, coils using 35-permeability powdered iron cores have been developed. In Table XVII are listed data for the coils now used in exchange area loading.

TABLE XVII

Coils for Loading Exchange Area Cables

Coil Code No.	Inductance (Henrys)	Core Permeability	Resistance-Ohms		Over-all Dimensions Inches	
			D-C.	1000 Cycles	Diameter	Height
602	0.088	35	8.9	10.5	3.6	1.3
603	0.135	35	12.8	14.1	3.6	1.3
574	0.175	60	4.6	10.6	4.5	2.1

Effective resistance values are for a line current of 0.001 ampere.

The standardization of the small size Nos. 602 and 603 loading coils has made it possible to design containing cases and assembly methods which permit much larger numbers of coils to be enclosed in cases conforming to the dimensional limitations set by existing vault conditions. A series of cases having capacities up to 300 coils has now been developed. The use of these large potting complements will be of considerable value in reducing the space congestion encountered in the "downtown" sections of the larger metropolitan areas.

In the 300-coil case, a total of 1,200 soldered joints are required to connect the coil terminals to the stub cable conductors. It was accordingly very important that the assembly method should involve a minimum liability to open circuits, crosses, or grounds. To accomplish this, a method was devised whereby the various spindles of coils were assembled to a skeleton frame to which the cable stub containing the 600 terminal pairs is also attached. All splices to the outgoing conductors are made immediately adjacent to the individual

coil terminals, after which the skeleton unit consisting of the coils and stub cable with case cover attached, is picked up with suitable tackle, and the coil unit inserted in the case. Fig. 17 illustrates this stage of the assembly. The case is subsequently filled with moisture-proof compound and sealed in the usual manner.



Fig. 17—Assembly of 300-coil case
Lowering loading coils into case after coil spindles have been mounted on frame and coil terminals spliced to stub cable conductors

Installation Features. In general, the exchange area cables on which loading is required are run in underground ducts and consequently the great bulk of the exchange area loading is installed in underground vaults. Fig. 18 shows a typical loading installation in a "double-deck" vault in New York City. The loading coil cases

are placed in the lower part of the vault permitting the coil terminal stub cables to be brought up vertically behind the horizontal cable runs and spliced to the trunk cables in such a way as to minimize the difficulties of future work on the cables passing through the vault. The trunk cables enter the vault through ducts which may be seen at

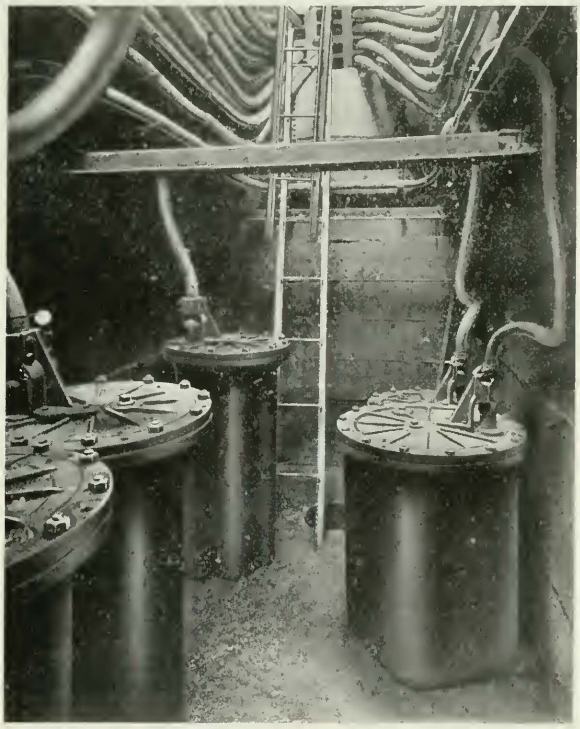


Fig. 18—Underground cable loading coil installation in Metropolitan area. Double-deck vault having ultimate capacity of 14 large coil cases

the top of the picture, and are supported on racks mounted on the upper side walls of the vault.

At present, a total of eight loading coil cases is installed in the vault illustrated in Fig. 17. Only five of these, however, appear in the picture. The cases now in place contain a total of 645 loading coils. The vault has space for six additional large cases, on which basis it is estimated that this vault will ultimately contain about 2,400 coils. Some of the largest vaults are capable of accommodating a total of 30 large cases containing a total of 9,000 coils.

VII. Loading for Submarine Cables

Coil Loading. The special problem of applying coil loading to submarine cables is a mechanical one, rather than one concerning the principles of loading. The situation in the United States is such that only a few coil loaded submarine cables have been required; this, of course, does not refer to the considerable number of instances where the submarine cables are so short that ordinary types of coils installed at the terminals satisfy the transmission requirements.

To date there have been installed in the United States a total of five cables having submarine coil loading; details of which are given in Table XVIII.

TABLE XVIII

Coil Loaded Submarine Cables

Location	Year of Installation	Length of Cable Miles	No. of Load Points	Number of Loaded Ccts.	Spacing of Coils-Miles	Coil Inductance Henrys
Chesapeake Bay No. 1	1910	4.5	2	17 pr. 13 A.w.g.	1.97	0.117
Chesapeake Bay No. 2	1916	4.0	1	12 qd. 13 A.w.g.	2.0	0.067-S 0.042-P.
Tarrytown-Nyack	1916	2.7	2	37 qd. 16 A.w.g.	0.89	0.250-S 0.155-P
Raritan Bay No. 1	1917	5.3	5	37 qd. 16 A.w.g.	0.91	0.250-S 0.155-P
Raritan Bay No. 2	1918	5.3	5	37 qd. 16 A.w.g.	0.89	0.250-S 0.155-P

The Raritan Bay cables each have 37 quads loaded at five points, 111 coils at each point, constituting the largest installation of submarine coil loading in the world. The depth of water in which these cables are installed is about 35 feet.

In each of the above instances, dry core paper cables were used. The design of the coil was made to fit the special designs required for the submarine loading pots. The case design was such as to furnish complete protection of the coils against moisture penetration and adequate mechanical strength for taking up the tension in the cable. In installing the cables, the procedure was to splice the loading coil cases into the cable while the cable was coiled on a barge, and then lay the cable and coils as a continuous operation. The service record of these

loaded submarine cables is excellent, thus demonstrating a satisfactory solution of the many difficult mechanical problems involved.

Continuous Loading. Another form of submarine cable loading, first put into practise by the Danish engineer, C. E. Krarup,[30] is to wind an iron wire or tape spirally around the copper conductor. This gives a continuous loading which has found important applications in the case of telephone and telegraph cables laid in deep water. So far as land cables are concerned, it has been found that continuous loading is uneconomical in comparison with coil loading. The only instances of continuous loading in the plant of the Bell System are the Florida-Cuba cables,[31] connecting Key West and Havana, which are the longest and most deeply submerged cables in use for telephonic communication in the world.

VIII. Extent of Commercial Application

The following data will assist in visualizing the practical importance of the developments which have been described in this paper.

In 1911, when Mr. Gherardi addressed this Institute on the subject of loading practise in this country, there were about 125,000 loading coils in service which loaded about 85,000 miles of open wire circuits and 170,000 miles of cable circuits. Although precise figures are not yet available regarding the number of loading coils in service in the Bell System as of January 1, 1926, conservative estimates set this total at about 1,250,000 coils. These coils load about 1,600,000 miles of cable circuits and 250,000 miles of open wire. In round numbers, 500,000 coils are installed on non-quadded local area trunk cables and 700,000 in toll and toll entrance cables (the bulk of these being quadded cables). Nearly two-thirds of the total number of coils have compressed iron powder cores, all of these being installed on cable circuits. About 4500 coils having wooden cores are installed on carrier loaded entrance cables. The remainder have iron wire cores, approximately 60,000 being of the so-called "air-gap" types.

Prior to the development of satisfactory types of telephone repeaters, the principal use of loading coils was in exchange area trunk cables in large metropolitan areas such as New York, Chicago, Philadelphia, and Boston. The successful application of telephone repeaters to loaded small gage cables has greatly increased the use of loading in the telephone plant. As illustrating this trend, approximately 150,000 toll

[30] C. E. Krarup, Submarine Telephone Cables with Increased Self-Induction, *ETZ.*, 23:344, April 17, 1902.

[31] W. H. Martin, G. A. Anderegg, B. W. Kendall, "Key West-Havana Submarine Telephone Cable System," Trans. A. I. E. E., Vol. 41, 1922, p. 184.

cable coils were manufactured for the Bell System in 1925, and approximately 100,000 exchange area cable coils. Recent estimates of the loading coil requirements for the next five years indicate an annual demand at a rate which would double the total number of loading coils in service about 1930.

As regards the field of application for cable loading in terms of cable lengths, the entrance and intermediate cables represent the minimum lengths; for instance, pieces as short as 500 feet when present in carrier telephone systems may require loading. In the local exchange areas, toll switching trunks as short as two miles may require loading. On the other hand, as illustrating the longest circuit now entirely in cable, a connection between Boston and Milwaukee—via New York, Pittsburgh, Cleveland, and Chicago—typifies the possibilities in the existing repeatered loaded cable plant. The over-all length of such a circuit is approximately 1200 miles. There is no technical obstacle to the use of repeatered loaded cables for distances several times as great; i.e., in the present state of the art, this is primarily a question of economics rather than of development.

IX. Conclusion

It will be appreciated from the foregoing account that the invention of coil loading was the beginning of an era of intensive development which has been marked by enormous advances in the design of telephone transmission lines, and that there has been no slackening of the inventional or development activity devoted to this subject. It is significant that at present more engineers and physicists in the departments represented by the authors are engaged on loading development problems than at any previous time.

In this account of the progress of the loading art during the past quarter century, the authors have endeavored to point out the relation of the loading developments to other phases of telephone development such as cables, repeaters, telegraph working, and carrier telephone and telegraph systems. In the space that is available, it would be impracticable to assign full credit to the many individuals who have been engaged in the development work on loading and the related problems. The final accomplishments should be regarded as the result of well coordinated efforts along many lines.

In conclusion, it may be of interest to note what the development and use of loading has meant to the telephone using public from an economic standpoint. Leaving out of consideration altogether loading on long toll cables—where the interdependence of repeaters and load-

ing is such that it is impracticable to assign to each its share of the savings—and taking into consideration only the loading of interoffice trunks and toll open wire circuits, it has been estimated that the larger wires which would have been required to give the present grade of transmission if loading had not been available, taken together with the heavier pole lines and additional underground ducts, would have entailed an additional investment in Bell System telephone plant of over $100,000,000.

BIBLIOGRAPHY

In addition to the references already cited, the following will be of interest:

Vaschy, *Annales Telegraphiques*, 1886, p. 321.

Heaviside, *Electrician* (London) June 3, 1887.

Vaschy, *La Lumiere Electrique*, 1889, 31, 83.

Heaviside, "Electromagnetic Theory," Vol. 1, 1893, p. 441.

S. P. Thompson, "Ocean Telephony," *Proceedings* Electrical Congress, World's Fair at Chicago, 1893.

M. I. Pupin, "Propagation of Long Electrical Waves," Trans. A. I. E. E., Vol. 16, 1899, 93.

W. A. J. O'Meara, "Submarine Cables for Long Distance Telephone Circuits," *Journal I. E. E.* London, April, 1910.

F. Bresig, *Theoretische Telegraphie*, 1910.

J. A. Fleming, "Propagation of Electric Currents," Van Nostrand Company, 1911.

A. E. Kennelly, "The Application of Hyperbolic Functions to Electrical Engineering Problems," University (London) Press, 1912.

A. Ebeling, *ETZ.*, 1914, p. 695 and p. 728.

K. W. Wagner, *Archiv für Electrotechnik*, Vol. 3, 1915, p. 315.

J. G. Hill, "Telephonic Transmission," London, Longmans, 1920.

K. W. Wagner, and K. Küffmüller, *Archiv für Electrotechnik*, Vol. 9, 1921, p. 461.

Sir Wm. Noble, "Long Distance Telephone System of the United Kingdom," *Journal I. E. E.*, Vol. 59, p. 389, 1921.

Das Fernsprechen Im Weitverkehr—Reichspostministerium, Berlin, 1923.

R. S. Hoyt, Impedance of Smooth Lines and Design of Simulating Networks, *Bell System Technical Journal*, April, 1923.

L. D. Cahen, Le Progres et l'Etat Actuel de la Technique des Liques Pupinisees—Bulletin de la Société francaise des Electriciens Aug.-Oct. 1924 (4 Serie), 4:755.

K. S. Johnson, Transmission Circuits for Telephonic Communication, Van Nostrand Co., 1925.

A Static Recorder

By H. T. FRIIS

Synopsis: This paper discusses different types of apparatus for recording static and also describes a new instrument in which the output of the set is kept constant by automatic control of the amplification, this amplification then being recorded as the relative measure of static. The set makes use of a fluxmeter with zero restoring torque by means of which the rectified output current arising from static interference is integrated over a period of ten seconds. The following five seconds are required to adjust the gain of the amplifier and record the change in gain from an arbitrary level. The gain is recorded in stops of 4 TU which correspond to a power amplification change by approximately a factor of 2.5. A record is shown during which the intensity of static changed by a factor of more than 10,000.

IN the following is given a general discussion of receiving sets for recording static and also a detailed description of a new instrument of this kind which is based upon the principle that the output of the set is kept constant by automatic control of the amplification, this amplification then being recorded as the relative measure of the static.

The literature on manual measurements of static is plentiful and for extensive references the reader may be referred to a paper on "Present Status of Atmospheric Disturbances," presented by L. W. Austin before the American Geophysical Union.[1] Many different methods of measuring static have been employed in obtaining the results given in this paper and it may further be added that we have found that the most reliable method of measuring the effect of static upon the intelligibility of speech signals is to introduce a local warbler signal [2] in the antenna. Unfortunately, however, all manual measurements require trained observers and therefore the cost of making continuous measurements will always be high, and besides, the human element introduced will decrease their reliability.

Very little has been published on automatic recording of static. The American Telephone and Telegraph Company and the Western Electric Company in 1923 developed an automatic static recorder which measured the high frequency currents induced in a loop antenna by amplifying them and passing them through a recording thermocouple meter. This apparatus was also equipped with a means of automatically measuring the gain of the entire receiving device so that the energy of the static could be evaluated directly. A popular account of this device was given under the caption "Getting Static's Autograph" by Austin Bailey in "Popular Radio," May, 1924. A recorder working at Aldershot, England, is mentioned in a paper by

[1] Will be published in the Proc. I. R. E. probably in the February, 1926, number.

[2] See "Radio Transmission Measurements," by Bown, Englund and Friis, Proc. I. R. E. Vol. 11, No. 2.

R. A. Watson Watt,[3] but it seems that this recorder is mainly an automatic counter of static crashes and it would therefore be of little value in U. S. A. where static is mostly a continuous rumble. The reason for the small advance which has been made to date in the automatic recording of static is probably due largely to the lack of suitable apparatus. Certainly there has never been any doubt that automatic records would be very valuable. It is just as important to know the static level as it is to know the strength of a radio signal because it is the static to signal ratio that determines the intelligibility of the signal. A static recorder connected to a rotating directional antenna system would tell us where static comes from and therefore enable the radio engineer to determine whether it is worth while to construct a directive antenna system. Also the connection between thunder-storm areas and static would make static recording valuable to the meteorological service. There is perhaps no reason why a suitable static recorder should not make it possible in a few years to obtain a daily static forecast just as we get our weather forecast now.

The question is then, what would be the best way of obtaining such a record of static? It would, of course, be very desirable to get a continuous record of the actual shape of the static wave, but we have no hope of ever realizing this and will have to be satisfied with the wave forms of a few typical static impulses as given by Watson Watt and E. V. Appleton.[4] Besides it would require a tremendous amount of labor to interpret such a record. The recorder described in this paper records the energy received within periods of 10 seconds. To be sure, such an energy curve of static does not tell the whole story due to the fact that the character of static is so variable. Thus, the same energy levels of a continuous rumbling static and of static consisting of separate clicks does not mean that these two types of static have the same effect upon the intelligibility of a speech signal. However, the shape of an energy record will indicate the general character of the static, but whether such an energy record will enable us to obtain absolute quantitative results with respect to the effect of static upon speech signals cannot yet be determined until further experimental results are available.

Requirements of an Energy Recorder

Let us take the case of recording the rectified current through the receiver of an ordinary receiving set supplied with a local carrier

[3] "Directional Observations of Atmospheric Disturbances," by R. A. Watson Watt, Proc. Royal Soc. A, Vol. 102, page 477.

[4] "On the Nature of Static," by Watson Watt and Appleton. Proc. Royal Soc. A, Vol. 103, page 84.

oscillator. This may, for instance, be accomplished by replacing the phone by a thermocouple connected to a standard recording galvanometer. Such a recorder would probably record the average of the current squared, but only for changes of energy received of not more than fifty times. The daily variation of the energy level of static (at 60 kilocycles) is, however, generally at least 100 times and sometimes even 10,000 times, so that this method of employing a receiving set with fixed amplification is unsatisfactory. Besides it would be very difficult to prevent overloading of the set if we limit ourselves to the use of 10-watt tubes.

One important requirement of a static recorder is therefore that the *output level of the receiving set be kept constant*, or more correctly, within

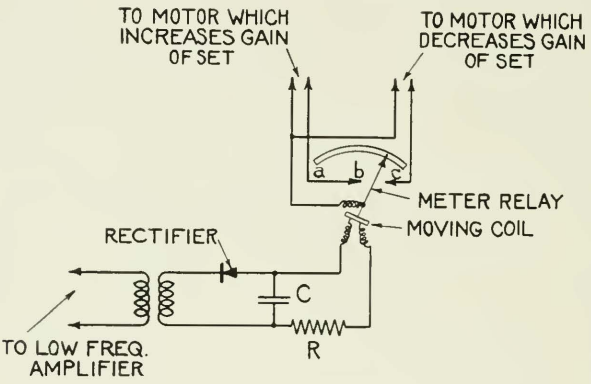


Fig. 1—Continuous recording system

certain narrow limits. In the receiver here described, the amplification of the receiving set is varied so as to satisfy this requirement, i.e., the gain is automatically cut down when the output level reaches its upper limit, and vice versa.

The output level of the set may be kept constant either by a continuously or by a discontinuously working system. The first method is mentioned here for comparison purposes only and is illustrated in Fig. 1. The figure shows that the rectified output current is sent through a moving coil relay whose pointer can only move between the contact points b and c while its real zero position is at a. Nothing happens if the pointer is between b and c but at the moment it touches b a motor will start, to increase the gain of the receiving set and will continue until the pointer is free again. Correspondingly the gain will be decreased if the pointer touches c. The purpose of the large resistance R and the condenser C is to prevent quick movements of

the pointer caused by the individual static crashes. The time constant *RC* of the circuit would probably have to be at least five seconds and the speed at which the gain of the set is changed must be correspondingly slow. This system will, therefore, react very slowly for great changes in static level. Ordinarily static does not change very fast but if the recorder is working with a directive antenna system that is rotated say 360° in 20 minutes, then a fairly fast working recorder is desirable. The main disadvantage of the continuous system is that it will be difficult to make the meter give a true indication of the average energy level.

The method employed is therefore based upon a discontinuous system and will be described in connection with Fig. 2. The rectified output current of the set is integrated over a period of 10 seconds by means of a fluxmeter.[5] If, after these ten seconds the fluxmeter deflection is below a certain mark, then the gain of the set is increased *one* step and, vice versa, if the deflection is above a certain mark the gain is decreased *one* step. For deflections in between these marks the gain remains unchanged. To change the gain one step and to bring the fluxmeter needle back to zero takes approximately 5 seconds after which the whole process is repeated. The output energy due to static received during ten-second periods is here kept within two *definite limits*. The gain can be changed only one step after each period, but since each step corresponds to a change of 4 *TU* (1.58 times) in voltage gain it will take only one minute and a quarter for the recorder to adjust itself to a sudden change of 100 times in the energy level of static.

The Apparatus of the Recorder

The receiving set is shown schematically in Fig. 2. It is an ordinary double detection set that requires altogether ten tubes, of which the last low frequency amplifier tube must be able to handle 10 watts in order to prevent overloading. The power supply may be rectified *AC*.

The gain control is inserted in the first intermediate frequency amplifier in order to be sure that no tubes are overloaded. The local oscillator shown is used for amplification calibration of the set and

[5] A galvanometer with negligible restoring torque, whose deflection is proportional to the coulombs sent through it. Its use for the present purpose was suggested by Mr. L. J. Sivian of the Bell Telephone Laboratories. He has employed the instrument for measurements of rectified speech and noise currents on telephone circuits. The use of a fluxmeter for similar purposes has been independently reported by Dr. E. M. Terry, of the University of Wisconsin, at the December 30, 1924, meeting of the U. R. S. I. at Washington, D. C.

requires no special shielding as its input voltage induced into the loop is comparatively large.

The selectivity of the set is determined by three separate units, viz., the antenna circuits, the intermediate frequency filter and the low frequency filter, each of which has a specific use. Carson and Zobel [6] have made the following statement.

"In filters designed to select a band of frequencies of width w, the ratio of energy transmitted through the network by the signal and by

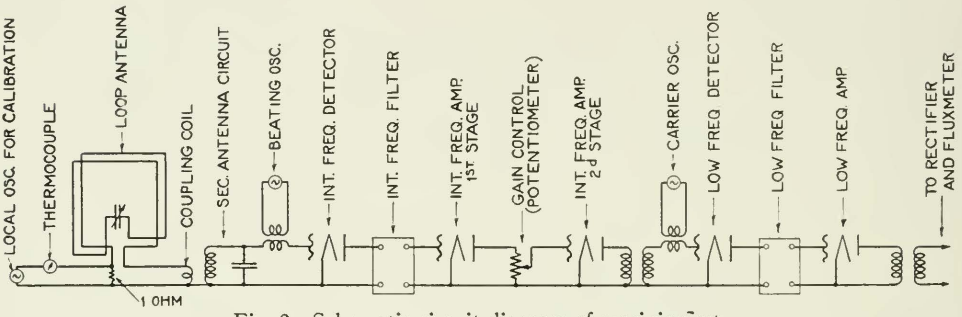


Fig. 2—Schematic circuit diagram of receiving set

random interference is inversely proportional to the band width and increases inappreciably when the number of sections is increased beyond two."

The main purpose of the filters is therefore not to define the frequency band of the set insofar as static is concerned, but to exclude continuous wave interference. It is hoped that 500 cycles wide frequency bands [7] can be maintained free of *c.w.* interference for static measurements and the simplest way to obtain such a band in the receiver is to make the low frequency filter an efficient low pass filter that cuts off every frequency above 600 cycles. More than two coupled circuits are hardly required in the antenna circuits, but the intermediate frequency filter ought to have sharper cut-off points than two coupled circuits will give. The selection of filters naturally depends upon the *c.w.* interference and it may in some cases be possible to reduce the number of filters and thereby make the recorder cheaper. The records shown later correspond to a frequency band of 2000 cycles—between 57.5 and 59.5 *k.c.*,—but it will probably not be long before *c.w.* interference makes it necessary to reduce this band

[6] "Transient Oscillators in Electric Wave-Filters"—John R. Carson and Otto J. Zobel, Bell System Technical Journal, Vol. II, No. 3, p. 27.

[7] Bands at 15, 30, 60, 120 . . . kilocycles would probably be satisfactory.

width. It is desirable to have a loud speaker connected to the output of the set and occasionally listen for *c.w.* interference.

The constant output control apparatus is shown in Fig. 3. The fluxmeter is seen in the upper right corner. Full deflection corresponds to 2×10^{-4} coulomb. The needle is normally free to move except when the cam Z presses the needle down until its point touches the scale OS. The shaft carrying the cam Z and the disc N is rotated

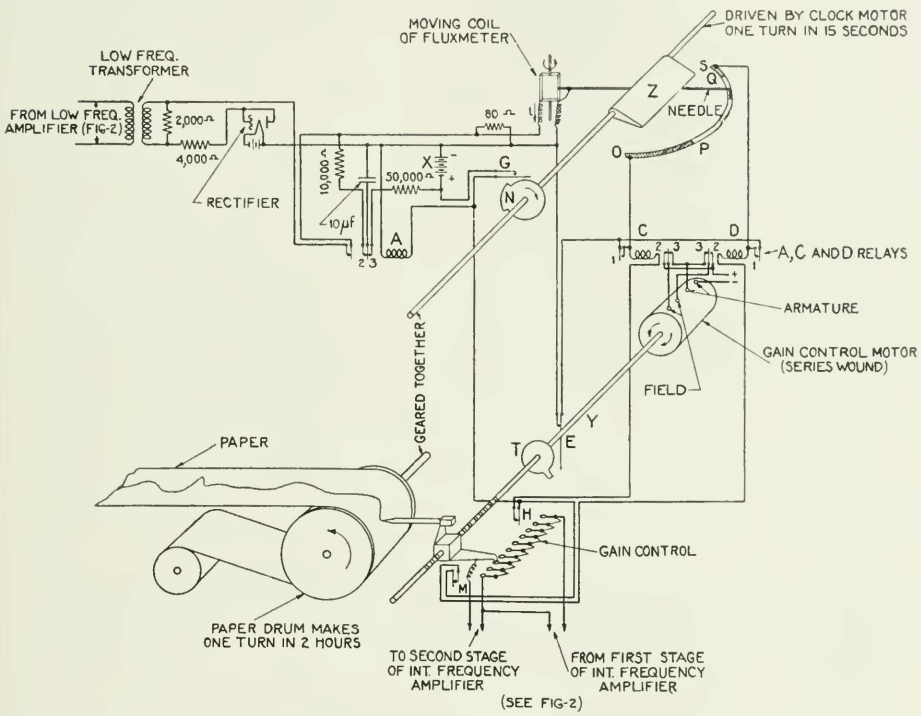


Fig. 3—Schematic circuit of constant output control apparatus

one complete turn in 15 seconds by a clock motor. The different elements are explained in the figure and the whole action may be understood by studying this carefully. However, it is probably worth while to go through a complete 15 second period and explain in detail the purpose of each part.

Time in Seconds—0–10:

Switch G is open, therefore relay A is open and no current can pass through the windings of relays C and D. (These relays start the gain control motor, which is therefore shut off.)

Contact 1 of relay A is closed and closes the circuit consisting of the secondary winding of the low frequency output transformer, the rectifier for the static currents

and the fluxmeter. The 2000 and 4000 ohm resistances in this circuit insure distortionless input voltage to the rectifier. The fluxmeter is damped by an 80 ohm shunt. The needle, which was initially at zero, will therefore move, its deflection being proportional to $\int i dt$.

Time in Seconds—10–14:

Switch G is closed by the cam on the revolving disc N and locks relay A.

Contact 1 of relay A is opened and opens the rectifier fluxmeter circuit, thereby bringing the fluxmeter needle to a stop.

Contact 3 of relay A is closed and makes the battery X charge the 10 μf condenser through the 50,000 ohm resistance.

Time in Seconds—11–14:

The cam Z presses the needle point down on the scale OS. Now, one of three things will happen.

1. Static has decreased since the last period, so that the needle point will make contact with the metal strip OP and close the following circuit: Battery X, needle of fluxmeter, winding of relay C, switch H and switch G to battery X. Relay C is therefore closed and its closed contact 2, together with contact 3 of the open relay D will start the gain control motor. After approximately half a turn of the gain control or motor shaft Y the needle point is lifted from OP by the rotation of the cam Z, but relay C stays closed due to the fact that it is self-locking through its contact 1, so that the shaft Y continues turning until the switch E is opened by the disc T. This opens the self-locking circuit of relay C. Relay C therefore opens and the gain control motor stops after the shaft Y has made exactly one complete turn and increased the gain of the set one step (4 Transmission Units or 1.58 times). Notice that the opening of the needle point contact does not break any current, due to the use of self-locking relays. This preserves the needle point contact.

2. Static has not changed since the last period. The needle point will now touch the insulating strip PQ and nothing else will happen, i.e., the gain of the set remains unchanged.

3. Static has increased since the last period so that the needle point now will make contact with the metal strip QS and close relay D and as in case 1 the motor will start and turn the shaft Y one turn, but this time in the opposite direction, i.e., the gain of the set is decreased one step.

Time in Seconds—14:

Switch G is opened again by the revolving disc N and opens relay A. Contact 2 of relay A is closed and will discharge the 10 μf condenser through the fluxmeter, thereby bringing the needle back to zero. (Notice that the time constant of this discharge circuit is $10000 \times 10 \times 10^{-6} = 1/10$ seconds.)

Time in Seconds—15:

A new period has started.

The purpose of the switches M and H is to stop the motor when the gain control switch arm has reached the end of the scale.

The recorder is of such recent development that no comprehensive data are yet available.

Fig. 4 shows part of an actual record of static received on a set tuned to 57.5—59.5 kilocycles. The ordinates represent the attenuation of the gain control of the set and it is to be remembered that the gain of the rest of the set is constant. The curve shows that the static power on the morning of October 30 changed more than 10,000 times. The point B on the curve gives the effect of inducing a local



Fig. 4—Static record, morning of Oct. 30, 1925, Cliffwood, N. J., U. S. A.

signal of strength 380 $\mu v/m$ in the loop.[8] The point C on the curve shows that at 8:25 A.M. the static intensity received on a 2000 cycle wide frequency band corresponded to the energy received from a *c.w.* signal of strength 3.8 $\mu v/m$. It would be practical always to relate static to such a *c.w.* signal. Experiments are now being conducted to determine whether the energy received from static is proportional to the width of the frequency band of the receiving set and if such is found to be the case then it is proposed to have the data relate to a 1000 cycle wide band. That static is, say, 7 microvolts per meter per kilocycle (7 $\mu v/mkc$) would then mean that the energy of the static received on a 1000 cycle wide frequency band is the same as the energy received from a *c.w.* signal of strength 7 $\mu v/m$.

Attempts have been made to calibrate the set by inducing in the loop, voltages of the shape shown in Fig. 5. Relating static to such signals would have the advantage of being independent of the band

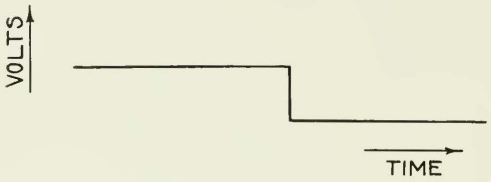


Fig. 5—Shape of impulse voltage

width of the set. Such signals were obtained by closing and opening a mercury switch, but one signal per second, or 10 impulses per period, would overload the set (the tubes) very much. At least 10 impulses per second would be required if the set should not be overloaded by each individual impulse, but this would be a difficult task to accomplish and it is therefore recommended that static be measured as explained above, by inducing a local *c.w.* singal into the loop. The fact that five static crashes in the course of 10 seconds—one period—does not overload the set while 100 impulses of the shape shown in Fig. 5 are required to prevent overloading gives us some interesting information on static. It shows that a single static crash is not a single sudden change of the field in the ether and that it cannot be represented by less than 20 consecutive impulses.

The record of Fig. 4 shows that each step on the gain control potentiometer is 4 TU and the selection of such steps and of 15 seconds will now be discussed. To decrease the 4 TU step to a 1 TU step would

[8] It may be worth while to have such a calibration signal introduced automatically for instance once every two hours.

decrease the speed of the set, i.e., it would take four times longer for the recorder to register a sudden change in the static level which is particularly a disadvantage when the recorder is connected to a rotating directional antenna. On the other hand a step larger than 4 *TU* would not give the static level with sufficient accuracy. If the time periods are changed from 15 to 10 seconds, then the "speed" of the set is increased, but the set is then inoperative over a larger part of the period since it takes 5 seconds to change the gain of the set and bring the fluxmeter needle back to zero. Besides, such a decrease in time period would increase the probability of overloading and also it would make the energy received per period vary more irregularly especially if static consisted of separate crashes.

Directional Static Recording

The usefulness of a static recorder will naturally be increased many times if it is able to indicate the general direction from which the static comes. A rotating loop antenna would give some results, but it would be still better to combine the rotating loop with an ordinary open antenna so as to obtain the well-known heart-shape directional characteristic. The loop should be rotated by the clock-motor of the set (see Fig. 3), say 2 complete turns in one hour, and the abscissa on the record would then require a direction scale in addition to the time scale.

Directive Diagrams of Antenna Arrays

By RONALD M. FOSTER

SYNOPSIS: Two systematic collections of directive amplitude diagrams are shown for arrays of 2 and of 16 identical antennae spaced at equal distances along a straight line with equal phase differences introduced between the currents in adjacent antennae, assuming that each antenna radiates equally in all directions in the plane of the diagram. Three diagrams show the effect of increasing without limit the number of antennae in a given interval. Two models show the effect of distributing the antennae over an area.

INTRODUCTION

ONE of the means proposed for obtaining directive radio effects, both in sending and in receiving, is the antenna array, consisting of a system of two or more antennae situated at specified fractions of a wave-length apart and with relations imposed upon the amplitudes and phases of the currents in the several antennae. For example, consider a sending array consisting of two vertical antennae so arranged that the currents in the antennae are equal in magnitude but a half period apart in phase, the individual antennae being identical and radiating equally in all directions in the horizontal plane. If the two antennae were placed at the same point there would be zero transmission in all directions, since the effects of the two antennae would neutralize each other. If, however, the two antennae are separated by a small fraction of a wave-length, while there will still be zero transmission in the direction perpendicular to the axis of the array, there will be transmission in all other directions. If this separation is increased to exactly one-half of a wave-length, the radiation from the array along the axis will become a maximum.

This particular type of antenna array was proposed by Brown [1] in 1899. A few years later, Stone [2] proposed a similar array with the two currents exactly in phase. This array gives maximum transmission perpendicular to the axis, zero transmission along the axis. About the same time, Blondel [3] made several suggestions, among them, two antennae placed a quarter of a wave-length apart and with a phase difference of a quarter of a period. With this arrangement a unilateral effect is obtained, there being maximum transmission in one direction along the axis, zero transmission in the opposite direction.

[1] S. G. Brown, British Patent No. 14,449 (1899).

[2] J. S. Stone, United States Patent No. 716,134 (1901).

[3] A. Blondel, Belgian Patent No. 163,516 (1902), British Patent No. 11,427 (1903).

These early suggestions have been followed by a number of more complicated arrangements proposed by Braun,[4] Bellini,[5] and others.[6]

Most books on radio communication contain one or more directive diagrams showing the variation of either the amplitude or the energy for systems of two, three, or four antennae which are separated by given fractions of the wave-length and with currents which have assigned amplitude and phase relations. Bellini,[7] Koerts,[8] and Zenneck [9] each give about a dozen such diagrams. Recently, Green [10] and Friis [11] have published directive diagrams for a pair of loops; the latter gives a curve obtained experimentally which agrees very well with the theoretical curve. Of the published diagrams, one of the most extended and systematic sets appears to have been that of Walter.[12] He showed a total of 21 diagrams for arrays of two antennae, with the three separations of 1/10, 1/4, and 1/2 wave-length and the seven phase differences of 0, 1/12, 1/6, 1/4, 1/3, 5/12, and 1/2 period.

In the present paper an effort has been made to present a more systematic and comprehensive collection of directive diagrams for arrays consisting of 2 and of 16 antennae, respectively, spaced at equal distances along a straight line or axis, with currents of equal amplitude in all the antennae, and with equal phase differences introduced between the currents in adjacent antennae. These diagrams are polar diagrams showing the relative amplitude of the field of the radiation at a great distance in a plane through the array, assuming that each antenna radiates equally in all directions in this plane. The unit circle shown in each diagram represents the amplitude of the radiation if all the antennae were made coincident in space and in phase.

These directive diagrams may be used to obtain the directive diagram in any plane through an array made up of antennae which

[4] F. Braun, Electrician, 57, pages 222–224, 244–248, 1906.

[5] E. Bellini, Electrician, 74, pages 352–354, 1914.

[6] For extensive bibliographies see L. H. Walter, Directive Wireless Telegraphy, London, 1921, pages 119–121; H. H. Beverage, C. W. Rice, and E. W. Kellogg, Journal of the A. I. E. E., 42, pages 736–738, 1923; A. Koerts, Atmosphärische Störungen in der drahtlosen Nachrichtenübermittlung, Berlin, 1924, pages 149, 150; J. Zenneck and H. Rukop, Drahtlose Telegraphie, fifth edition, Stuttgart, 1925, Chapter XIII.

[7] E. Bellini, Jahrbuch der drahtlosen Telegraphie und Telephonie, 2, pages 381–396, 1909.

[8] A. Koerts, *loc. cit.*, pages 101, 102, 104, 105, 110, 111, 130, 131, 133.

[9] J. Zenneck and H. Rukop, *loc. cit.*, pages 412–415, 419, 421, 423, 428, 432.

[10] E. Green, Experimental Wireless, 2, pages 828–837, 1925.

[11] H. T. Friis, Proceedings of the I. R. E., 13, pages 685–707, 1925.

[12] L. H. Walter, Electrician, 64, pages 790–792, 1910.

do not radiate equally in all directions in this plane, but which satisfy the other conditions named above; the total directive effect is the product of the individual effect multiplied by the group effect. Thus, since the amplitude of the radiation in the horizontal plane from a single vertical loop varies as the cosine of the angle between the direction of transmission and the plane of the loop, the directive diagram in the horizontal plane of an array of loops, all loops being oriented the same, is the corresponding directive diagram of an antenna array as presented in this paper, with the radius vector multiplied by a cosine factor.

The present discussion has been stated in terms of transmission, but the directive diagrams apply equally well to the case of reception by an array from a distant source.

Each of the two sets of diagrams is presented in a rectangular arrangement so as to exhibit the effect of changes both in the separation between adjacent antennae, specified in wave-lengths, and in the phase difference introduced between the currents in adjacent antennae, specified in periods. These drawings were originally made at the suggestion of Dr. G. A. Campbell to illustrate the application of antenna arrays as a means for reducing the ratio of static to signal.

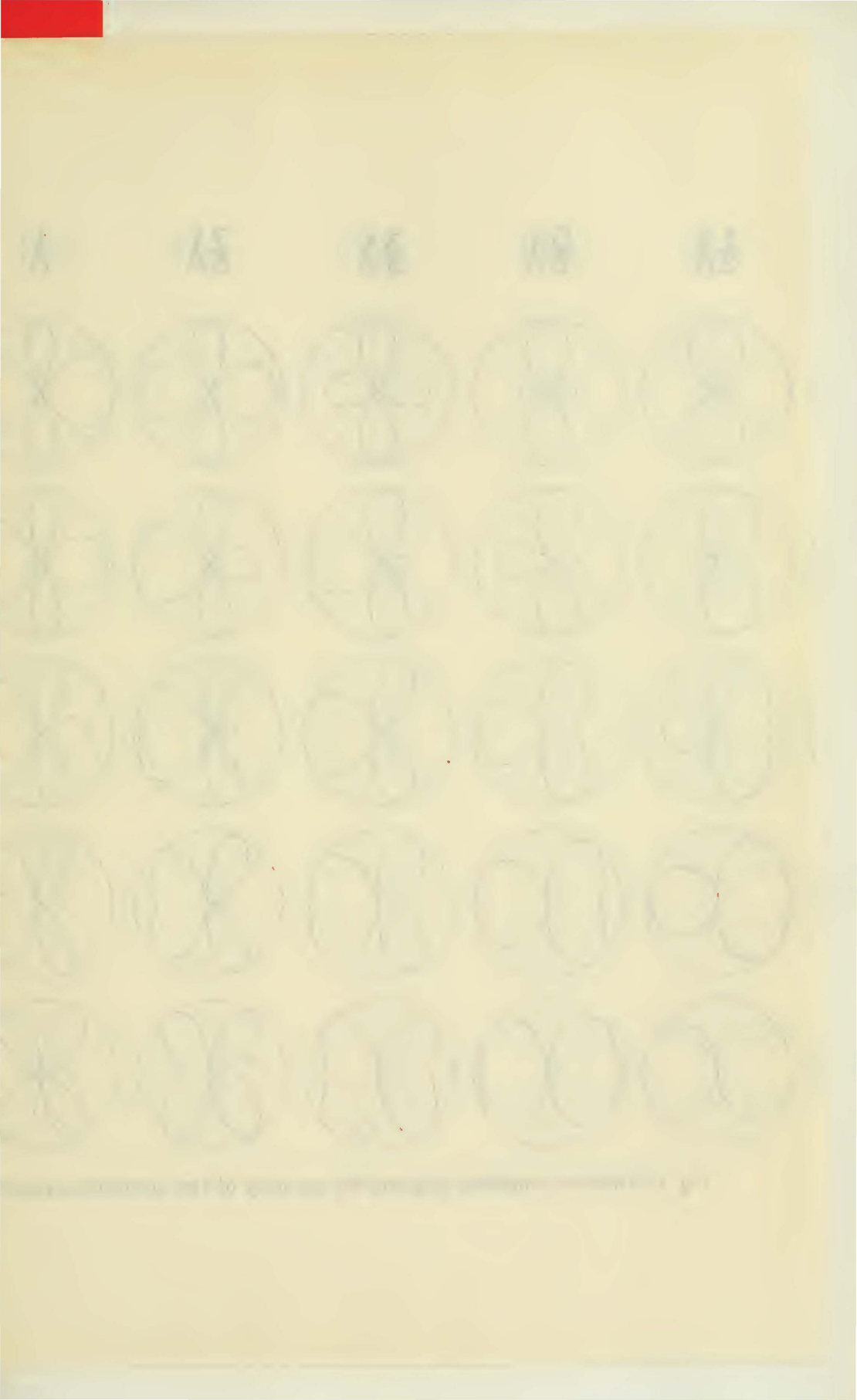
The antenna array is analogous to the optical diffraction grating. By eliminating the transmission wires connecting together the individual antennae of the array and utilizing instead re-radiation from suitably designed antennae, the radio system would correspond more closely with this optical analogue. With this arrangement, however, the phase difference cannot exceed a value in periods numerically equal to the separation in wave-lengths, a restriction to which the ordinary ruled grating is also subject. The retardation grating proposed by Rayleigh [13] and the echelon spectroscope of Michelson [14] offer more complete analogies to the antenna array in that there is no theoretical limitation on the separation and phase difference.

Two Antennae, Fig. 1

A total of 90 directive diagrams for an array of two antennae is shown by Fig. 1. The separation between antennae varies from 0 to 2 wave-lengths, in steps of 1/8 wave-length; the phase difference between antennae varies from 0 to 1/2 period, in steps of 1/8 period; an additional set of diagrams is included with a separation of 4 wave-lengths. These curves were carefully drawn with a unit circle ten

[13] Rayleigh, Collected Papers, 3, pages 106–116.

[14] A. A. Michelson, Astrophysical Journal, 8, page 37, 1898.



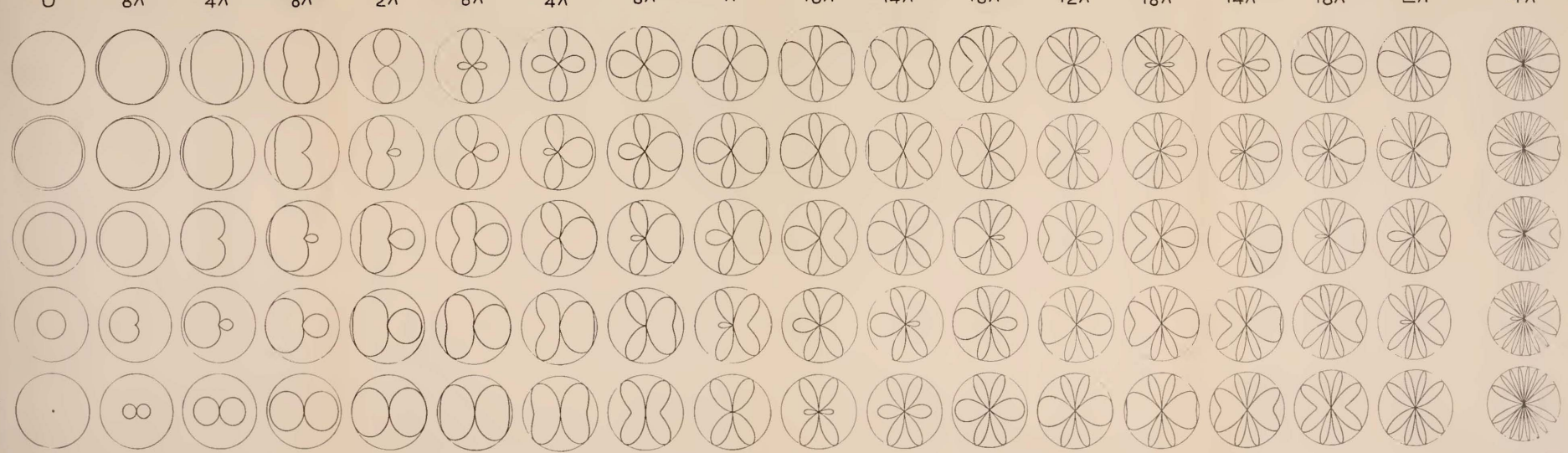


Fig. 1—Directive amplitude diagrams for an array of two antennae; separation in wave-lengths (λ) along the top, phase difference in periods (T) at the left

inches in diameter, so that, on the reduced scale of reproduction, the accuracy should leave nothing to be desired. For a sending array the specified phase difference is the lag of the current in the right-hand antenna behind the current in the left-hand antenna; for a receiving array it is the lag introduced in the current from the right-hand antenna. In each case the line of the array is parallel to the horizontal axis of the diagram.

Reversing the sign of the phase difference reflects the directive diagram about its vertical axis, that is, the right and left sides are interchanged. With increasing phase difference the diagrams repeat cyclically, those from $\frac{1}{2}T$ to $\frac{3}{2}T$ being the same as those from $-\frac{1}{2}T$ to $\frac{1}{2}T$, and so on.

In the first column, that is, for zero separation, all diagrams are circles, since the two antennae are coincident, and thus the array radiates uniformly in all directions. For zero phase difference the directive diagram is the unit circle, since the radiations from the two antennae reenforce each other without interference. As the phase difference is increased, this circle grows smaller, due to increasing interference, until, for a phase difference of a half period, the two radiations completely neutralize each other, the directive diagram shrinking down to a null circle.

The diagrams in the first row, that is, for zero phase difference, are symmetrical about the vertical axis in addition to being symmetrical about the horizontal axis. In every case the amplitude is unity along the vertical axis, that is, in a direction perpendicular to the line of the array. As the separation is increased from zero, the amplitude along the horizontal axis diminishes, until it reaches zero for a separation of $\frac{1}{2}\lambda$, it then increases to unity at λ, it diminishes to zero at $1\frac{1}{2}\lambda$, it reaches unity at 2λ, and so on.

The diagrams in the bottom row, that is, for a phase difference of a half period, are also symmetrical about the vertical axis. In every case the amplitude is zero along the vertical axis. For small separations, the directive diagram is approximately a pair of tangent circles, which increase in size as the separation is increased. When the separation reaches $\frac{1}{2}\lambda$, the amplitude along the horizontal axis reaches unity, it then falls off to zero as the separation is increased to λ, it rises to unity at $1\frac{1}{2}\lambda$, it falls to zero at 2λ, and so on.

The diagram for $(\frac{1}{4}\lambda, \frac{1}{4}T)$ is particularly interesting in that there is a single direction of unit amplitude with zero amplitude in the opposite direction. This array was proposed by Blondel, as stated above, and it is the basis of the Alexanderson barrage.[15] The diagrams

[15] E. F. W. Alexanderson, Proceedings of the I. R. E., 7, pages 363–378, 1919.

inches in diameter, so that, on the reduced scale of reproduction, the accuracy should leave nothing to be desired. For a sending array the specified phase difference is the lag of the current in the right-hand antenna behind the current in the left-hand antenna; for a receiving array it is the lag introduced in the current from the right-hand antenna. In each case the line of the array is parallel to the horizontal axis of the diagram.

Reversing the sign of the phase difference reflects the directive diagram about its vertical axis, that is, the right and left sides are interchanged. With increasing phase difference the diagrams repeat cyclically, those from $\frac{1}{2}T$ to $\frac{3}{2}T$ being the same as those from $-\frac{1}{2}T$ to $\frac{1}{2}T$, and so on.

In the first column, that is, for zero separation, all diagrams are circles, since the two antennae are coincident, and thus the array radiates uniformly in all directions. For zero phase difference the directive diagram is the unit circle, since the radiations from the two antennae reenforce each other without interference. As the phase difference is increased, this circle grows smaller, due to increasing interference, until, for a phase difference of a half period, the two radiations completely neutralize each other, the directive diagram shrinking down to a null circle.

The diagrams in the first row, that is, for zero phase difference, are symmetrical about the vertical axis in addition to being symmetrical about the horizontal axis. In every case the amplitude is unity along the vertical axis, that is, in a direction perpendicular to the line of the array. As the separation is increased from zero, the amplitude along the horizontal axis diminishes, until it reaches zero for a separation of $\frac{1}{2}\lambda$, it then increases to unity at λ, it diminishes to zero at $1\frac{1}{2}\lambda$, it reaches unity at 2λ, and so on.

The diagrams in the bottom row, that is, for a phase difference of a half period, are also symmetrical about the vertical axis. In every case the amplitude is zero along the vertical axis. For small separations, the directive diagram is approximately a pair of tangent circles, which increase in size as the separation is increased. When the separation reaches $\frac{1}{2}\lambda$, the amplitude along the horizontal axis reaches unity, it then falls off to zero as the separation is increased to λ, it rises to unity at $1\frac{1}{2}\lambda$, it falls to zero at 2λ, and so on.

The diagram for $(\frac{1}{4}\lambda, \frac{1}{4}T)$ is particularly interesting in that there is a single direction of unit amplitude with zero amplitude in the opposite direction. This array was proposed by Blondel, as stated above, and it is the basis of the Alexanderson barrage.[15] The diagrams

[15] E. F. W. Alexanderson, Proceedings of the I. R. E., 7, pages 363–378, 1919.

situated on the line from $(0\lambda, \frac{1}{2}T)$ to $(\frac{1}{4}\lambda, \frac{1}{4}T)$ have a similar property: each has a relative maximum in a single direction, with zero in the opposite direction. The diagrams on the line from $(0\lambda, 0T)$ to $(\frac{1}{2}\lambda, \frac{1}{2}T)$ have a maximum along the horizontal axis to the left, with an amplitude to the right decreasing from unity at $0T$ to zero at $\frac{1}{4}T$, and then increasing to unity at $\frac{1}{2}T$.

The number of lobes tends to increase as the separation is increased, as shown by (a) of Fig. 2. A zigzag starting at $(0\lambda, \frac{1}{2}T)$

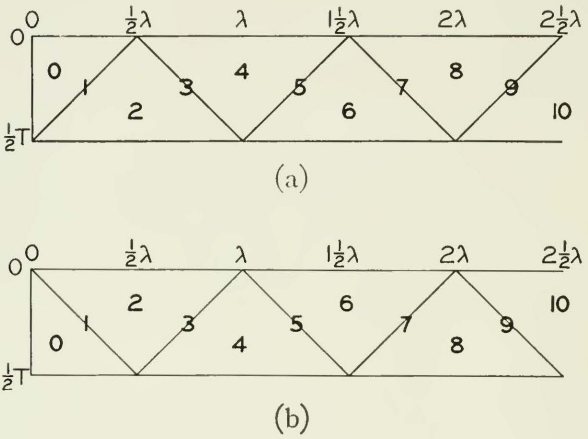


Fig. 2—(a) Number of null directions (which is also the number of lobes) for two antennae. (b) Number of unit directions (directions of absolute maximum amplitude) for any number of antennae, in terms of separation and phase difference between adjacent antennae

and made up of lines sloping up and down at an angle of 45° divides the rectangular arrangement of diagrams into sections with 0, 2, 4, 6, . . . null directions in each diagram, respectively. On these lines the number of null directions is 1, 3, 5, 7, . . . , respectively, with the intermediate numbers at the junction points. The number of lobes is, of course, equal to the number of null directions.

Part (b) of Fig. 2 is a diagram specifying the number of unit directions (directions of absolute maximum amplitude) in terms of the separation and the phase difference between adjacent antennae, and it holds regardless of the number of antennae, that is, the number and position of the main lobes are not changed by increasing the number of antennae, provided the same separation and phase difference are preserved between adjacent antennae.

It is interesting to observe the variation in the diagrams along any line in this rectangular arrangement of Fig. 1, whether horizontal,

vertical, or diagonal. A lobe starts as a small bud, it grows in size until it reaches the unit circle, it then becomes dented; the two prongs of the lobe separate more and more until a division into two lobes takes place; then these lobes separate as a new lobe starts to grow

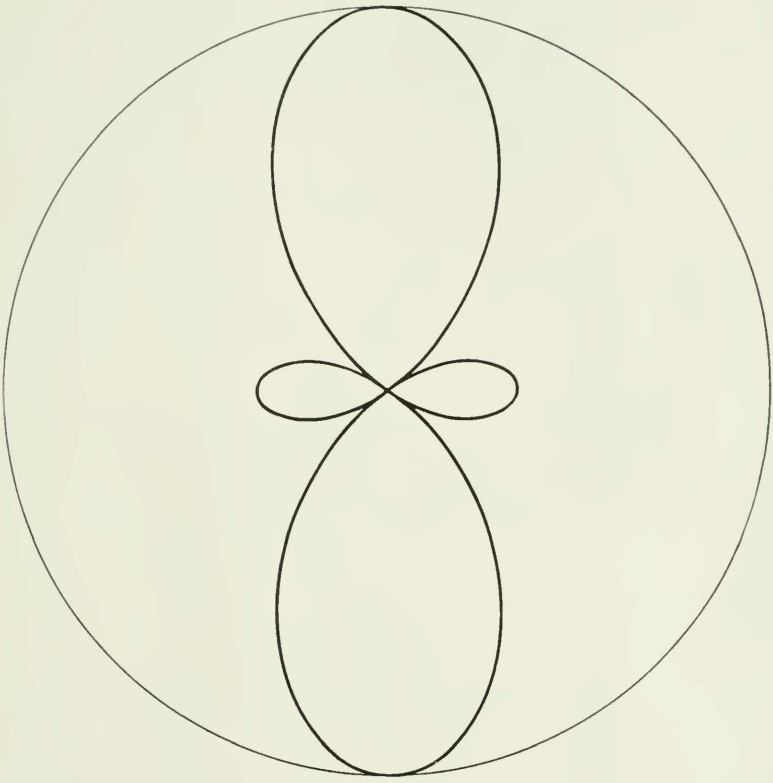


Fig. 3—Directive amplitude diagram for two antennae (0.6098λ, 0T) having the minimum area (0.2986) relative to the unit circle

between them. The additional column of diagrams for a separation of 4λ still further illustrates the way in which the lobes multiply and narrow as the separation between the two antennae in increased.

The area of the polar directive diagram of an array relative to the area of the unit circle is a measure of the reduction in the energy ratio of random static to signal for that array, assuming that the signal comes from a direction in which the radius vector of the diagram is unity while the static is uniformly distributed. All the diagrams for a phase difference of 1/4 period have an area of 1/2.

The area of the other diagrams oscillates about 1/2 and approaches it as a limit upon increasing the separation and keeping the phase difference constant. The minimum area (for a diagram in which the radius vector reaches its maximum of unity [16]) is 0.2986, obtained

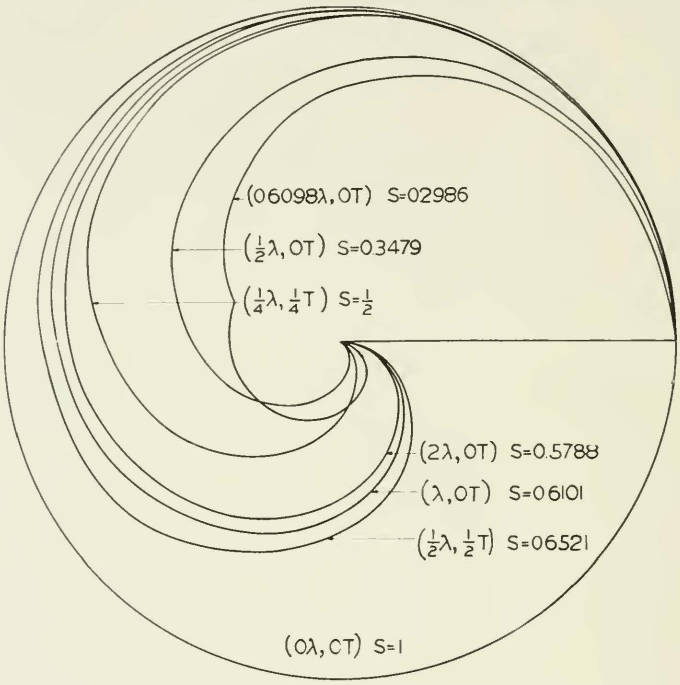


Fig. 4—Cumulative amplitude diagrams for two antennae

by the array (0.6098λ, 0T). The directive diagram for this case is shown by Fig. 3.

Cumulative amplitude diagrams are shown by Fig. 4 for six selected arrays, including the unit circle, which is the cumulative diagram for the array (0λ, 0T). In this figure, the angle corresponding to any value of the radius vector is equal to the total angle of the directive diagram of the array throughout which the relative amplitude

[16] A. Koerts, *loc. cit.*, pages 104, 105. In those cases in which the radius vector does not reach the unit circle, in order to obtain a measure of the reduction of the energy ratio of random static to signal, the unit circle should be replaced by the circle with a radius equal to the maximum radius vector. The absolute minimum 0.2986 is not changed upon including these cases but a relative minimum 1/3 occurs for the array ($a\lambda$, bT) upon letting a and b approach 0 and 1/2 in such a manner that $a+2b=1$. If the two antennae are loops, with planes parallel to the axis of the array, the area approaches the relative minimum 3/14 upon letting a and b approach the same limits 0 and 1/2 but in such a manner that $3a+4b=2$.

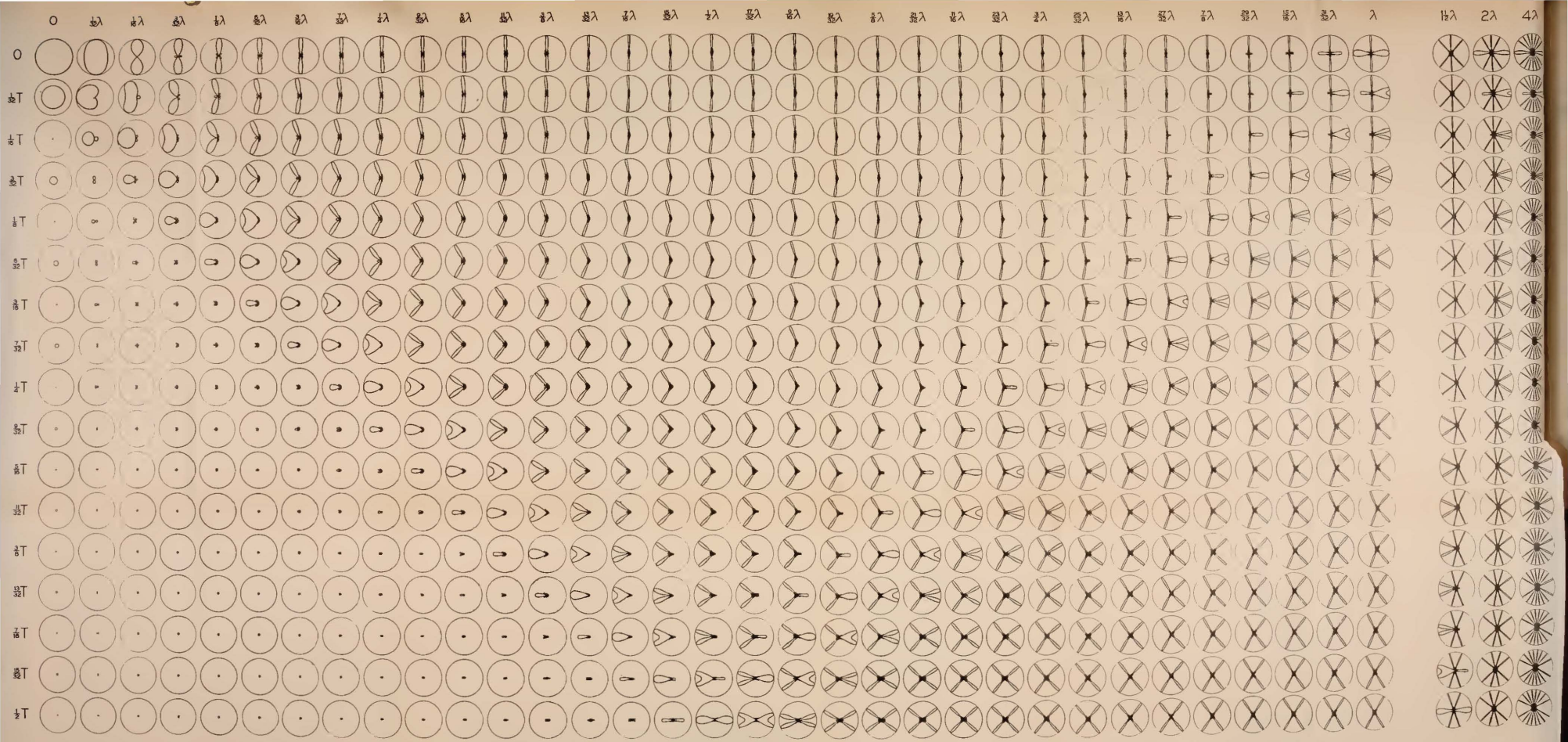


Fig 5—Directive amplitude diagrams for an array of sixteen antennae; separation in wave-lengths (λ) along the top, phase difference in periods (T) at the left

is equal to or greater than this value. The area (S) relative to the unit circle is given in Fig. 4 for each of the cumulative diagrams, this area being equal to the area of the corresponding directive diagram.

Sixteen Antennae, Fig. 5

A total of 612 directive diagrams for an array consisting of sixteen antennae is shown by Fig. 5. The separation between adjacent

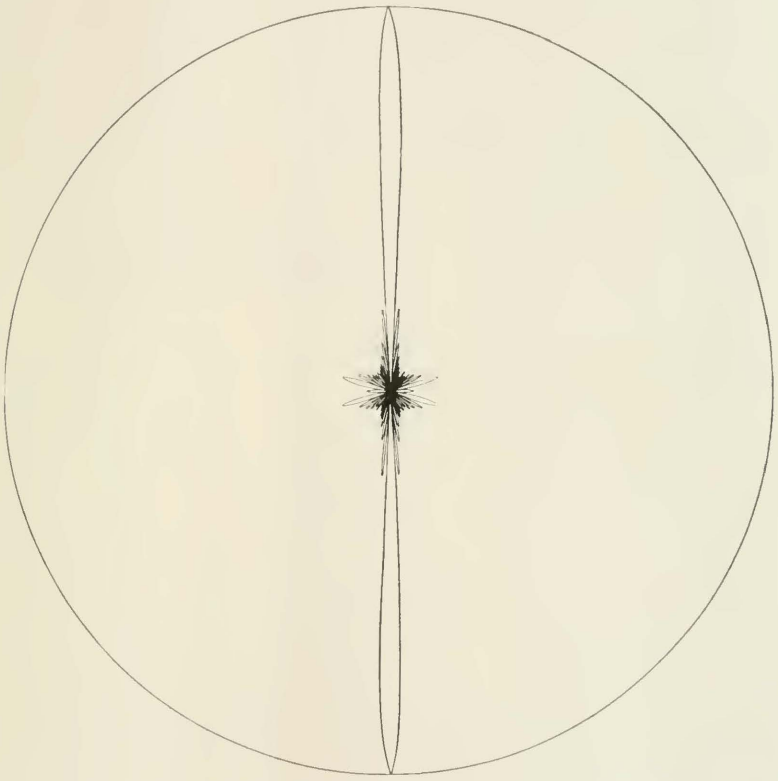


Fig. 6—Directive amplitude diagram for sixteen antennae (0.8825λ, 0T) having the minimum area (0.0254) relative to the unit circle

antennae varies from 0 to 1 wave-length, in steps of 1/32 wave-length; the phase difference between adjacent antennae varies from 0 to 1/2 period, in steps of 1/32 period; additional sets of diagrams are included with separations of $1\frac{1}{2}$, 2, and 4 wave-lengths. The specified phase difference is the lag of the current in one antenna behind the current in its left-hand neighbor. The diagrams are reflected about the vertical axis upon changing the sign of the phase difference, and they

is equal to or greater than this value. The area (S) relative to the unit circle is given in Fig. 4 for each of the cumulative diagrams, this area being equal to the area of the corresponding directive diagram.

Sixteen Antennae, Fig. 5

A total of 612 directive diagrams for an array consisting of sixteen antennae is shown by Fig. 5. The separation between adjacent

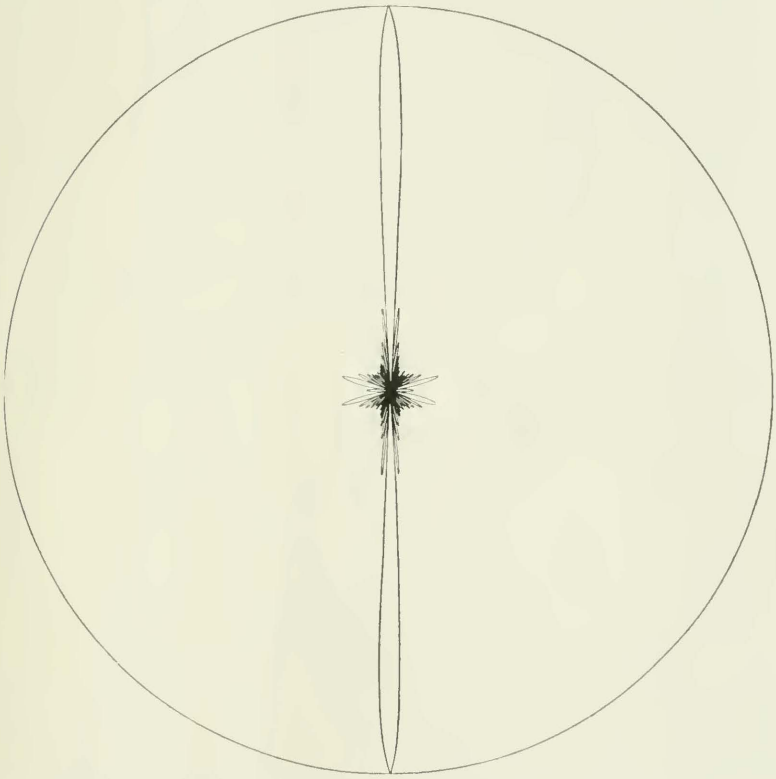


Fig. 6—Directive amplitude diagram for sixteen antennae (0.8825λ, 0T) having the minimum area (0.0254) relative to the unit circle

antennae varies from 0 to 1 wave-length, in steps of 1/32 wave-length; the phase difference between adjacent antennae varies from 0 to 1/2 period, in steps of 1/32 period; additional sets of diagrams are included with separations of $1\frac{1}{2}$, 2, and 4 wave-lengths. The specified phase difference is the lag of the current in one antenna behind the current in its left-hand neighbor. The diagrams are reflected about the vertical axis upon changing the sign of the phase difference, and they

repeat cyclically with increasing phase difference. These curves were copied from the original drawing; detailed accuracy is not claimed, but the arrangement and relative sizes of the lobes are approximately correct.

Comparison of Figs. 1 and 5 shows that, for the same set of parameters, the main features of the two diagrams are similar, that is,

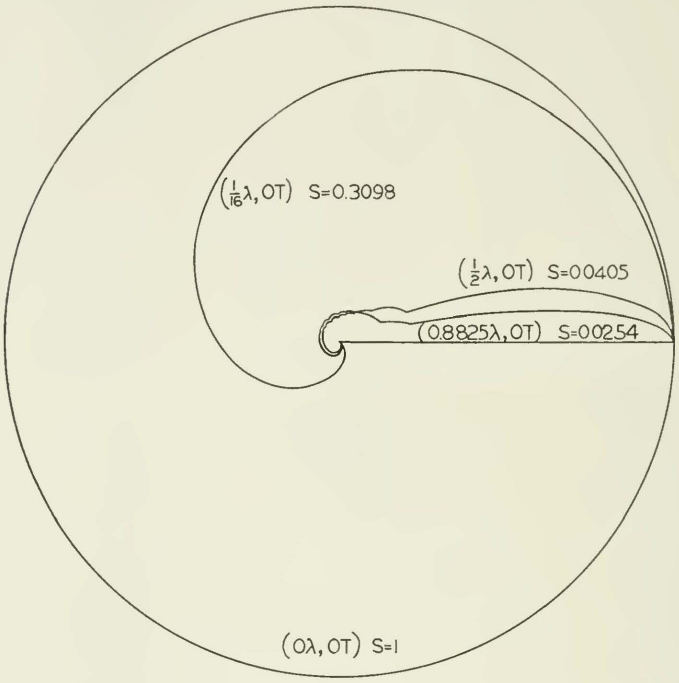


Fig. 7—Cumulative amplitude diagrams for sixteen antennae

the main lobes are located in the same positions, but with 16 antennae the main lobes are much narrower and in addition a multiplicity of small lobes occurs in many cases.

The area of the directive diagrams for 16 antennae oscillates about 1/16 and approaches it as a limit upon increasing the separation, keeping the phase difference constant. The minimum area (for a diagram in which the radius vector reaches its maximum of unity) is 0.0254, obtained by the array (0.8825λ, 0T). The directive diagram for this case is shown by Fig. 6. Cumulative amplitude diagrams for 16 antennae are shown by Fig. 7 for three selected arrays in addition to the array (0λ, 0T), the area of each diagram being given on the drawing.

Infinite Antenna Arrays

In view of the points of similarity between the diagrams for 2 and for 16 antennae, in particular for those pairs of diagrams for which the parameters of Fig. 1 are eight times the parameters of Fig. 5, provided the latter are relatively small, the question naturally arises as to the effect of increasing the number of antennae without limit.

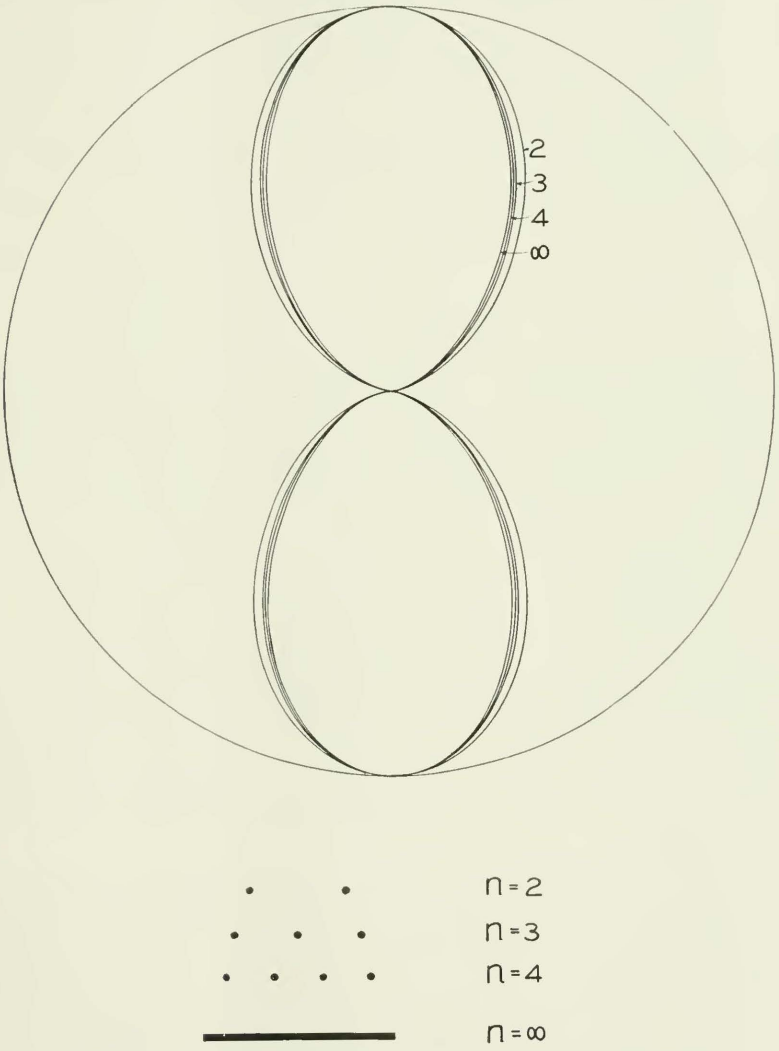


Fig. 8—Comparison of finite and infinite arrays within a total distance of one wavelength, with zero phase difference.

The similarity among the diagrams for arrays of antennae within a given fixed interval is illustrated by Figs. 8, 9, and 10.

In Fig. 8 are shown diagrams for 2, 3, and 4 antennae situated within a total distance of one wave-length with separations of 1/2, 1/3, and 1/4 wave-length, respectively, and with zero phase difference between adjacent antennae. The curve (∞) gives the limit of the family of directive diagrams for arrays of n antennae with a separation of $1/n$ wave-length, as n becomes infinite.

Fig. 9 gives similar curves for arrays of n antennae within an interval of two wave-lengths, with the parameters $\left(\frac{2}{n}\lambda,\ 0T\right)$. Fig. 10 shows a similar set of diagrams for arrays within an interval of one wave-length and within a total phase interval of one period, that is, for arrays of n antennae with the parameters $\left(\frac{1}{n}\lambda,\ \frac{1}{n}T\right)$.

For any interval $(A\lambda,\ BT)$ a similar family of curves can be obtained for arrays of n antennae with the parameters $\left(\frac{A}{n}\lambda,\ \frac{B}{n}T\right)$. As the number n is increased without limit, the directive diagram approaches a limiting curve. This limiting curve never has more than two directions of unit amplitude. There are zero, one, or two such directions, depending upon whether A is less than, equal to, or greater than B. The diagrams for the infinite case are reflected about the vertical axis upon changing the sign of the phase difference, but they do not repeat cyclically with increasing phase difference.

The rapidity with which the diagrams approach this limiting curve as n is increased is well illustrated by Figs. 8, 9, and 10. On the scale of these drawings, the curves for 16 antennae would be indistinguishable from the limiting curves for the infinite case. The upper left-hand corner of Fig. 5 may thus serve as a chart of the directive diagrams for the infinite case if the column and row headings are multiplied by the factor 16 to give the total separation and phase difference of the interval. For larger values of these parameters, however, the curves for the infinite case depart more and more from those of Fig. 5.

The diagrams with $A = B$ are of particular interest since these are unilateral, with the main lobe growing narrower as the total separation and phase difference are increased. In the case of the Beverage antenna, the ideal system [17] consists essentially of a long loop, which we may think of as the limiting case of a succession of a large number of narrow loops. The directive diagram of such an antenna system

[17] H. H. Beverage, C. W. Rice, and E. W. Kellogg, *loc. cit.*, pages 372, 373.

would, therefore, be the product of the group curve for an infinite number of antennae in the given interval multiplied by a cosine factor for the individual narrow loop.

Space Characteristics

When the antennae are not confined to a straight line but are distributed over an area, a surface with its radius vector propor-

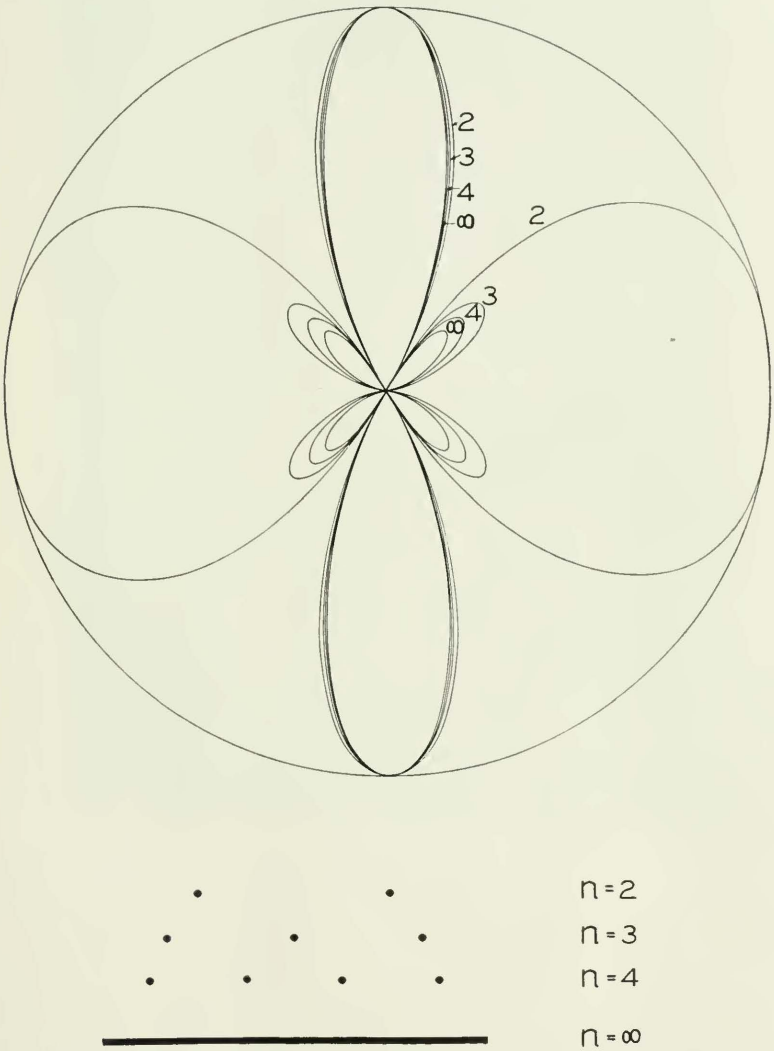


Fig. 9—Comparison of finite and infinite arrays within a total distance of two wavelengths, with zero phase difference

tional to the amplitude of the field of the radiation at a great distance from the array in the direction of the radius vector is required. Two particular cases will be illustrated in order to give some idea of the surface which shows the group effect of the array; for actual antennae, the radius vector of this surface must be multiplied by the corresponding radius vector of the space characteristic of the individual antenna in order to obtain the actual characteristic of the array.

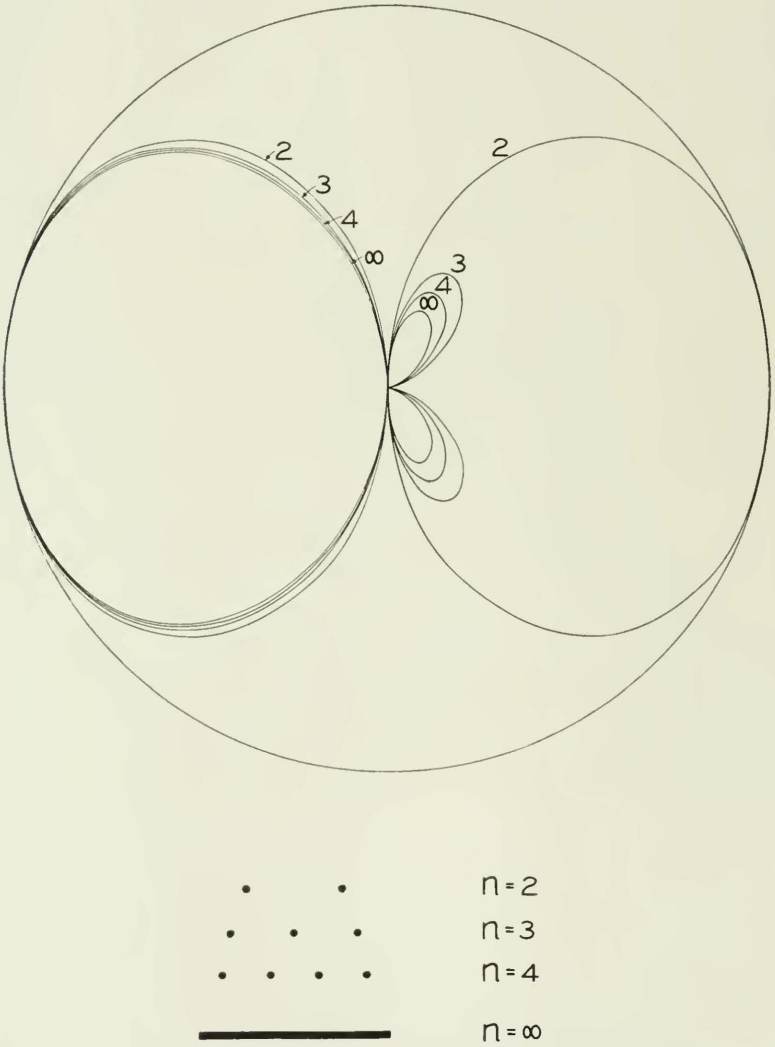


Fig. 10—Comparison of finite and infinite arrays within a total distance of one wave-length, and within a total phase interval of one period

A model of the upper half of the space characteristic of an array of four antennae located at the corners of a square is shown by Fig. 11, each side of the square having the parameters $(\frac{1}{2}\lambda, 0T)$. Below the model is shown the directive diagram in the plane of the array, which

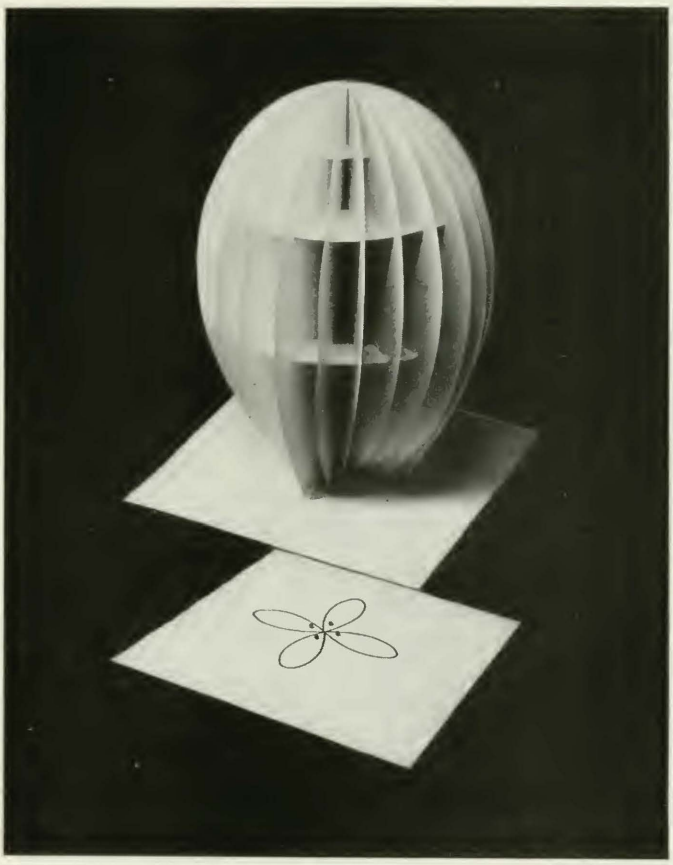


Fig. 11—Model of the space characteristic for an array of four antennae located at the corners of a square, with a separation of one-half wave-length between antennae on each side of the square, and with zero phase difference

is identical with the base of the model, together with a representation of the array itself. In the horizontal plane the maximum amplitude is slightly less than 1/5, occurring along the diagonals of the square; the amplitude reaches its absolute maximum of unity only in the vertical direction.

Fig. 12 shows a model of the upper half of the space characteristic of an array of 32 antennae located along the diagonals of a square, with the parameters ($\frac{1}{2}\lambda$, 07) in each diagonal. This space characteristic is a complicated surface, the main features of which are shown by the model, the smaller lobes not being shown clearly in detail. In

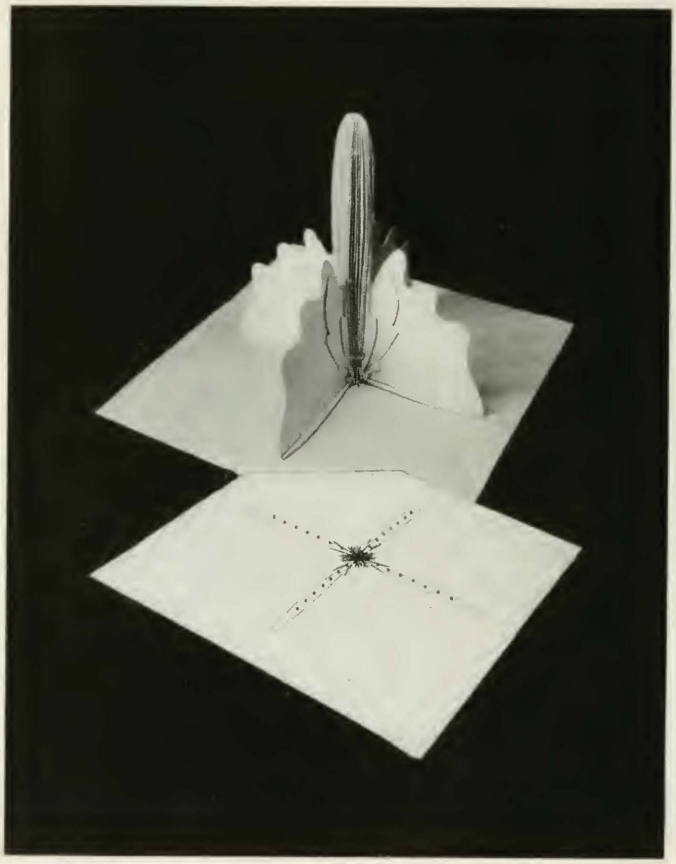


Fig. 12—Model of the space characteristic for an array of 32 antennae located along the diagonals of a square, with a separation of one-half wave-length between adjacent antennae in each diagonal, and with zero phase difference

the horizontal plane the maximum amplitude is 1/2, occurring along the diagonals of the square; the amplitude reaches a pronounced maximum of unity, however, in the vertical direction.

I am greatly indebted to Dr. Louisa E. Townshend for supervising the preparation of the drawings and models, and expecially for the accuracy attained in redrawing Fig. 1.

Appendix

Formulae for the directive diagrams of this paper are conveniently expressed in terms of polar coordinates, as follows: For a linear array of n antennae with the parameters $(a\lambda, bT)$,

$$r = \left| \frac{\sin n\ (\pi a \cos\theta + \pi b)}{n \sin\ (\pi a \cos\theta + \pi b)} \right|, \tag{1}$$

where θ is measured from the axis of the array. The area of this diagram, relative to the unit circle, is

$$S = \frac{2}{n^2}\left(\frac{n}{2} + \sum_{k=1}^{n-1} (n-k) J_0(2\pi k a) \cos\ (2\pi k b) \right). \tag{2}$$

For the special case $n=2$, Fig. 1, formula (1) reduces to

$$r = |\cos\ (\pi a \cos\theta + \pi b)|, \tag{3}$$

and formula (2) for the area to

$$S = \tfrac{1}{2}(1 + J_0(2\pi a) \cos\ (2\pi b)). \tag{4}$$

For the special case $n=\infty$ in a total interval $(A\lambda, BT)$, the limit of formula (1) for $a=A/n$ and $b=B/n$, as n becomes infinite, is

$$r = \left| \frac{\sin\ (\pi A \cos\theta + \pi B)}{\pi A \cos\theta + \pi B} \right|. \tag{5}$$

For the array of Fig. 11,

$$r = |\cos\ (\tfrac{1}{2}\pi \cos\theta \cos\phi) \cos\ (\tfrac{1}{2}\pi \sin\theta \cos\phi)|, \tag{6}$$

where ϕ is the angle which the radius vector makes with the plane of the array, and θ the angle which the projection of the radius vector in this plane makes with one side of the square. For the array of Fig. 12,

$$r = \left| \frac{\sin\ (8\pi \cos\theta \cos\phi)}{32 \sin\ (\tfrac{1}{2}\pi \cos\theta \cos\phi)} + \frac{\sin\ (8\pi \sin\theta \cos\phi)}{32 \sin\ (\tfrac{1}{2}\pi \sin\theta \cos\phi)} \right|, \tag{7}$$

where ϕ and θ are the same as for formula (6) except that the latter is measured from one diagonal of the square.

Correction of Data for Errors of Averages Obtained from Small Samples

By W. A. SHEWHART

SYNOPSIS: Recent contributions to the theory of statistics make possible the calculation of the error of the average of a small sample—something that cannot be done accurately with customary error theory. Obviously, these contributions are of very general importance, because experimental and engineering sciences alike rest upon averages which in a majority of cases are determined from small samples, and because an average cannot be used to advantage without its probable error being known.

The present paper attempts to show in a simple way why we cannot use customary error theory to calculate the error of the average of a small sample and to show what we should use instead. The points of interest are illustrated with actual data taken for this purpose. The paper closes with applications of the theory to four types of problems involving samples of small size for each of which numerous examples arise in practice. These types are:

1. Determination of error of average.
2. Determination of error of average difference.
3. Determination of most probable value of the root mean square deviation of the universe when only one sample of n pieces has been examined.
4. Determination of most probable value of the root mean square deviation of the universe when several samples of n pieces each have been examined.

USEFUL THEORY OVERLOOKED: WHY?

PRACTICALLY everyone uses averages—research workers and engineers in particular. Moreover, all of us have long appreciated the fact that an average is often only of value when we know its probable error. Naturally, we turn to the theory of errors to guide us in calculating the probable error. Naturally, because from 1733 to 1908 there was nothing else that we could turn to. Since 1908 the recognition has been gradually making headway that to use customary error theory for determining the probable errors of averages of small samples is a mistake.

The story of how to calculate the probable error of a small sample was originally told in Biometrika, a journal for the statistical study of biological problems—a veritable mine of useful information. The truth was given in equations involving terms familiar only to statisticians and hence was concealed from many. The story, however, with the aid of such experimental results as are used in this paper can be told in a simple manner: it is of interest to all of us who, for one reason or another, cannot make large numbers of observations on every quantity that we measure, but must nevertheless estimate the probable errors of our results. In this discussion, diagrams will be

used instead of equations, and, because of this rather popular presentation, many readers may want to consult, as the original sources, the intensely interesting mathematical contributions of "Student",[1] Professor Karl Pearson,[2] and R. A. Fisher.[3]

Case Where Customary Theory Applies

We start, as in customary error theory, with the assumption that the probability distribution of errors is normal. This simply means that the probability of the occurrence of an error within any range is assumed to be equal to the area under the so-called normal curve[4] (such a curve is shown in Fig. 1) between the limits of the same range.

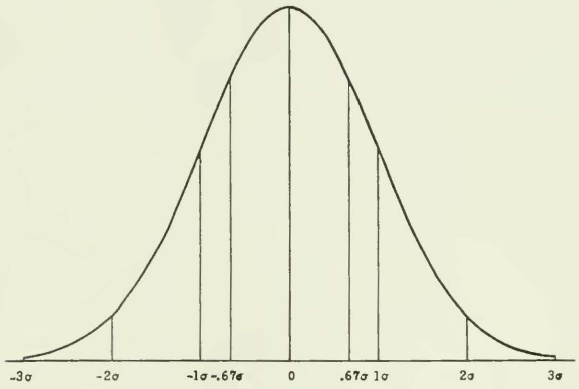


Fig. 1—Customarily assumed law of error curve—normal law

50.00000% of area within $0 \pm .67449\sigma$
68.26894% of area within $0 \pm 1\sigma$
95.44998% of area within $0 \pm 2\sigma$
99.73002% of area within $0 \pm 3\sigma$

The total area under the curve is, of course, unity. This curve is plotted with the origin at the true value and with the errors measured in units of the root mean square error σ. The fractions of the area bounded by certain multiples of the root mean square error are shown for reference.

Let us make an experiment and see how far customary error theory

[1] Biometrika, Vol. VI, 1908, pp. 1–15. Vol. XI, 1917, pp. 416–417.

[2] Biometrika, Vol. X, 1915, pp. 522–529.

[3] Biometrika, Vol. X, 1915, pp. 507–521. Proc. Camb. Phil. Soc., Vol. XXI, 1923, pp. 655–658.

[4] The equation for this has recently been traced back to Abraham De Moivre (1733) by Professor Pearson. See Biometrika, Vol. XVI, 1924, pp. 402–404.

carries us, see where it breaks down, see why it breaks down, and then avail ourselves of the new theory—a powerful tool of great value, because it makes possible for the first time the solution of many practical problems. Here is the experiment. Take 998 small circular chips, 499 green and 499 white. Mark 20 white ones with 0, 40 white ones with 0.1, 39 white ones with 0.2, etc., in accordance with the normal law. Do the same for the green chips except that all numbers on the chips are minus. Put the 998 chips in a bowl, mix thoroughly, draw out one and record it. Replace the chip, again mix thoroughly, and repeat the process until 4000 values are observed. A little reflection shows that this experiment is equivalent to making 4000 measurements of a quantity by a method subject to a normal law of error with a root mean square error of approximately unity.

Let us group these 4000 values into 1000 groups of 4, and determine the average for each group, taking the first four observations as the first group, the second four as the second group and so on. This gives

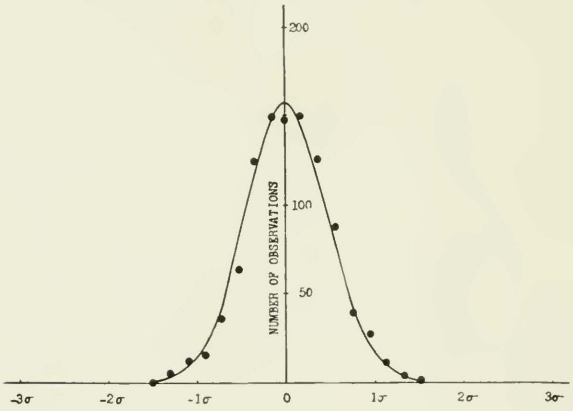


Fig. 2—Curve showing customary error theory to be satisfactory on one condition not often met in practice; *i.e.*, σ is known

• Distribution of 1000 averages of 4

– Normal law with root mean square error $\frac{\sigma}{\sqrt{4}}$

us 1000 averages. Suppose we subtract the true value m (in this case zero) from each average and divide this result by the root mean square error of the frequency distribution of values within the bowl. This gives us 1000 observations of the error of the average of 4 observations measured in terms of σ. Customary error theory shows that these averages should be distributed normally as indicated by the smooth

curve in Fig. 2 with a root mean square error of $\frac{\sigma}{\sqrt{4}}$ or one half that in Fig. 1. The dots show the experimental results.[5]

So far the customary error theory is satisfactory. But we do not often have this case in practice; that is, we do not know the root mean square error σ, and instead know only the observed root mean square error s of the sample.[6]

Case Where Customary Theory Does not Apply

Let us next recall just the way we use the customary theory in practice and then see what mistake we usually make. Take the results of drawing the first sample of 4 in the experiment previously cited. The four observed values are .6, −.2, 1.1, −2.0, the average $\overline{X}$ of these

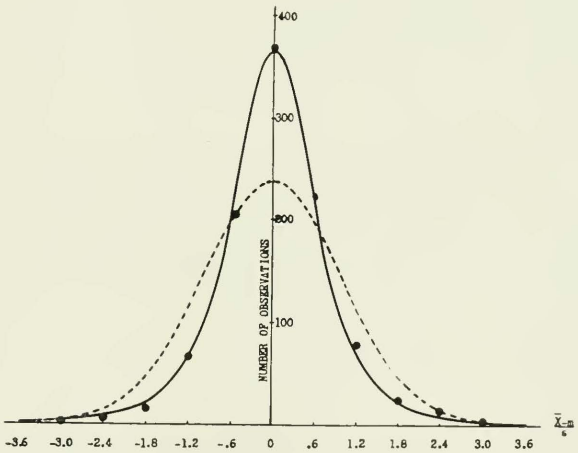


Fig. 3—Curves showing inaccuracy of customary error theory in finding error of average in terms of the observed standard deviation s

– – – – – Customary theory
——— New theory
• Distribution of 1000 z's

is −.125, and the observed root mean square deviation s is 1.177. Assuming no knowledge of the root mean square error σ of the distribution from which the sample of 4 was taken and using customary theory, we should assume the probable or 50% error to be $.6745\frac{1.177}{\sqrt{4}}$.

[5] I am indebted to Miss Victoria Mial and Miss Marion Cater for securing the experimental results, making all necessary calculations, and drawing the curves given in this paper.

[6] Customarily we do not know the true value m, hence instead of knowing the root mean square errors we know the root mean square or standard deviations.

This follows from the fact that the observed values of the ratio $z = \frac{\bar{X} - m}{s}$ where m is the true value, are customarily assumed to be distributed normally. *Here we come to the crux of the discussion: these observed values of the ratio are not distributed normally.* "Student"[7], in 1908, was the first to show how they are distributed.

Let us look at the observed frequency distribution of the 1000 z's given by the above experiment (dots Fig. 3). To be normally distributed, as customarily assumed, these dots would have to lie on the dotted normal curve. Obviously they do not. Instead they lie on a much more peaked curve (solid line) than the normal. This was calculated with the aid of "Student's" theory. We must therefore conclude: the probability that the mean of a sample of n, drawn at random from a normal distribution, will not exceed (in the algebraic sense) the mean of that distribution by more than z times the root mean square deviation of the sample cannot be found from the normal law when n is small. We must use the tables provided by "Student" in the two papers referred to above.

Why the Customary Theory Fails to Give the Error of the Average in Case of Small Samples

Let us look a little further into the reason why the z's are not distributed normally, before we consider the question as to the magnitude

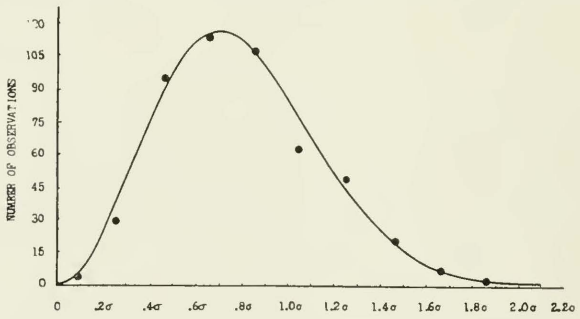


Fig. 4—Data furnishing a clue to reason for inadequacy of customary error theory

• Observed distribution of standard deviations of 1000 samples of four
– Theoretical curve of asymmetrical type

of the difference between the probable error determined from one theory and that determined from the other.

Let us look at the distribution of the 1000 standard deviations, the s's, Fig. 4, for here we shall find the secret revealed: The distri-

[7] Loc. cit.

bution of s's, as we might expect, is asymmetrical; the most probable standard deviation s, to be observed is not the average s. Of course, the customary theory assumes that the average s is the most probable s, and that the distribution of s is normal. We should therefore expect to find the z's distributed normally for values of n such that the dis-

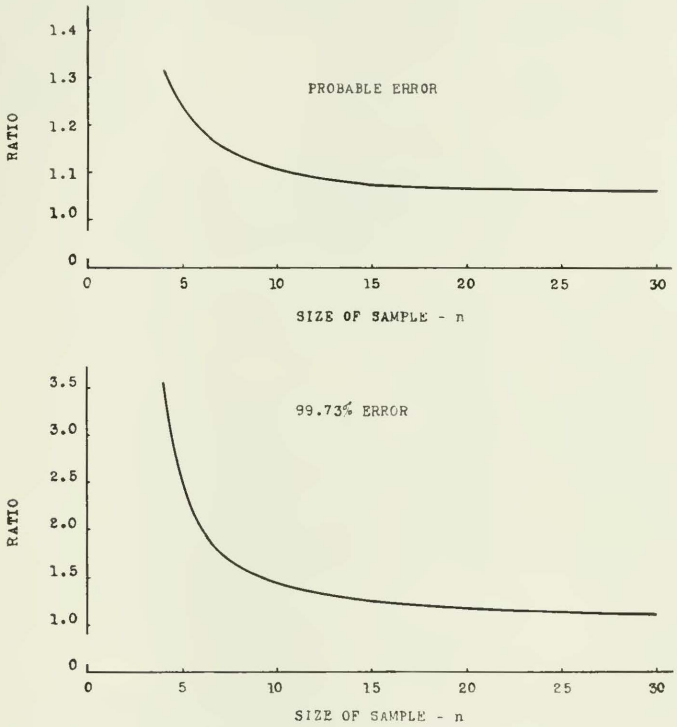


Fig. 5—Chart showing magnitude of correction for size of sample—ratio of the errors to their customarily accepted values

tribution of observed standard deviations is approximately normal. Now, Professor Pearson[8] has developed the theory underlying the distribution of s. He finds that as n increases, the distribution of s rapidly approaches normality. Even for n greater than 25 the distribution has approached normality to such an extent that we should expect the z's to be distributed approximately in normal fashion. The study of the distribution of z shows this to be true, as we shall see below.

In passing, we should note how closely the theoretical curve, Fig. 4, fits the observed points and also note two other checks between theory

[8] Loc. cit.

and observation furnished by the new data given herein. According to theory, the modal and mean values of s for samples of size 4 expressed in units of σ should be .707 and .798 respectively. The experimental results are .717 and .801.

How Much Larger Are the Probable and 99.73% Errors of an Average Than the Customarily Accepted Values?

The difference between the error of an average and its customarily accepted value increases as the number of observations n (or size of

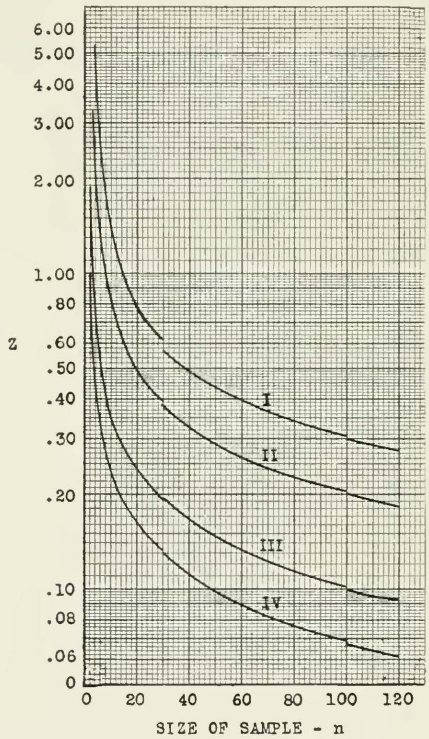


Fig. 6—Errors of averages of samples of size n
I -99.73002% error
II -95.44998% error
III -68.26894% error
IV -50.00000% error
z = the ratio of the error of the average to the observed standard deviation

sample) decreases. This fact is illustrated in Fig. 5. This figure shows the ratios of the errors to their customarily accepted values plotted for values of n from 4 to 30.

Curves showing the most frequently used errors of averages measured in terms of z (*i.e.* in terms of the ratio of the error to the observed standard deviation) are given in Fig. 6. The error curves for n less than 30 have been obtained with the aid of "Student's" original tables, those for n between 30 and 100 have been obtained from the normal law integral tables using the standard deviation of z; *i.e.* $\frac{s}{\sqrt{n-3}}$ as given by "Student." For n greater than 100, customary error theory has been used.[9]

Typical Practical Applications

But few, if any, recent developments of statistical theory are of more general application in most fields of scientific research and engineering than the one herein described.[10] This follows because the theory herein discussed must be used in calculating the required probable error (or other measure of dispersion) of the averages obtained from small numbers of observations. The number of applications of this character is legion.

Problem Type 1, Determination of Error of Average

Example 1:

Five samples of granular carbon taken from a crucible show resistances of 47.5, 49.4, 43.2, 48.0 and 46.2 ohms respectively. What are the probable and 99.73% errors of the average of these resistances?

Solution:

The observed values of average resistance $\bar{X}$, and standard deviation $s = \sqrt{\frac{\Sigma(\bar{X}-X)^2}{n}}$ are 46.9 ohms and 2.097 ohms respectively. Hence from Fig. 6 we see that the probable and 99.73% errors are respectively $.372s = .780$ ohms and $3.33s = .699$ ohms respectively whereas from customary theory they would be $.302s = .633$ ohms and $1.34s = 270$

[9] For the curves in this figure as in the preceding one, I have assumed the customary theory for the case where the true value of X is known so that the root mean square error of the average $\bar{X}$ of sample of size n is the ratio $\frac{s}{\sqrt{n}}$. Of course, as we know from customary error theory, if we assume no knowledge of the true value of X, we should use $\frac{s}{\sqrt{n-1}}$.

[10] Since this paper was written, a very interesting article, "Statistics in Administration," has appeared in *Nature* (V. 117, pp. 37–38, Jan. 9, 1926), calling attention to the importance of the theory of small samples.

ohms respectively. The true probable and 99.73% errors are 23% and 148% higher respectively than those calculated by customary theory, as is evident from Fig. 5.

Discussion of Type 1:

Examples of this type of problem are obviously so numerous that further illustrations need not be given. They occur every day in practically every science. We see that in such cases it is certainly necessary to allow for the effect of the small size of sample.

PROBLEM TYPE 2, DETERMINATION OF ERROR OF AVERAGE DIFFERENCE

Example 1:

Five instruments are measured for some characteristic X, first on one machine and then on another, giving two sets of values X_{11}, X_{12}, $\ldots X_{15}$, and X_{21}, X_{22}, $\ldots x_{25}$ respectively. Calculate the 5 differences $X_{11}-X_{21}=x_1$, $X_{12}-X_{22}=x_2$, $\ldots X_{15}-X_{25}=x_5$. Assume that the average difference is $\bar{x}$ and the standard deviation of the differences is s. Assuming the two machines give the same results except for random variations, what is the probability that the observed difference would occur? Are we justified in the assumption that the machines give the same results?

Solution:

The true difference is zero on this assumption. The observed difference is $z=\dfrac{\overline{X}-0}{s}$, and "Student's" tables may be used to evaluate this probability.[11] If this probability is very small, let us say .001 or less, it may be taken as indicating that the machines do not give the same results.

Example 2:

We wish to compare the depth of penetration obtained from two different methods of preserving chestnut telephone poles. We choose n poles for test. A sample from each pole is treated by one process, and a sample from each pole is treated by another process. The depths of penetration are measured. Are we justified in assuming the two methods to give significantly different results?

[11] Approximate values can be obtained from the curves in Fig. 6.

Solution:

If n is small, we proceed as in the previous case, to find the probability of occurrence of the observed difference. If this probability is small, we conclude that the difference is significant; *i.e.*, the two methods of preservation give different results.

Example 3:

Three-bolt guy clamps are used for clamping the guy wires on telephone poles. These are supplied from different sources. Those from one source fail to hold the wire as well as those from another and inspection shows that these same clamps fail to meet a certain specified dimension. The force required to slip the wire in each of 10 clamps from this source is measured. These clamps are then modified to meet the specified dimension and the force required to slip the wire in each clamp is again measured. Are we justified in attributing the failure to hold the wire to the fact that these clamps did not meet the specification?

Solution:

The solution follows the same line as in the first case.

Discussion of Problems of Type 2:

Problems of this type are very numerous. It is obvious that significant differences calculated as above indicated are always larger than those calculated by customary theory.

Problem Type 3, Determination of Most Probable Value of the Root Mean Square Deviation of the Universe When Only One Sample of n Pieces Has Been Examined

Example 1:

Five tool-made models are tested for their efficiency, giving values $X_1, X_2, \ldots X_5$. What is the most probable value of the range within which the efficiencies of product instruments may be expected to lie approximately 99.7% of the time, assuming that a manufacturing process can be developed which is the same as that used in producing the tool-made models?

Solution:

Customary practice would answer: the average of the five values plus or minus 3 times their standard deviation. The better answer is: the average plus or minus $\frac{3}{.7746}$ times the standard deviation.

This follows from Professor Pearson's work previously quoted. He has shown that the most probable observed standard deviation $\breve{s}$ of

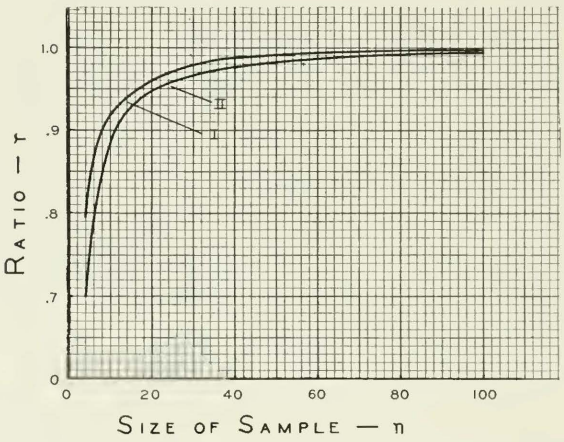


Fig. 7—Curves giving the most probable value of the true standard deviation σ
I When the average $\bar{s}$ of standard deviations of many samples is known. $r\sigma = \bar{s}$
II When the standard deviations of one sample is known. $r\sigma = \breve{s}$

a sample of n from a normal distribution with standard deviation σ is $\breve{s} = \sqrt{\frac{n-2}{n}}\,\sigma$. Substituting the value $n = 5$ in this equation we get $\breve{s} = .7746\sigma$.

A curve of the values of $\frac{\breve{s}}{\sigma}$ vs. n is presented in Fig. 7 for reference in solving problems of this character.

Problem Type 4, Determination of Most Probable Root Mean Square Deviation of the Universe When Several Samples of n Pieces Each Have Been Examined

Example 1:

One thousand transmitters, known to have different efficiencies, have been tested five times each for efficiency. Find the standard deviation of the machine method of measurement.

Solution:

Calculate the standard deviation of the five tests for each transmitter. Find the average value of these 1000 values and divide it by .8407. This follows from the fact that the average $\bar{s}$ of the observed standard deviation for a series of samples of size n drawn from a normal distribution with standard deviation σ is

$$\bar{s} = \sqrt{\frac{2}{n}} \frac{\left|\frac{n-2}{2}\right.}{\left|\frac{n-3}{2}\right.} \sigma.$$

where the symbol $\left|\underline{X}\right.$ is equivalent to $\Gamma\,(X+1)$

Thus for $n=5$ we get [12] $\bar{s}=.8407\sigma$.

Fig. 7 also presents the values of the ratio $\frac{\bar{s}}{\sigma}$ for reference and with sufficient accuracy for solving problems similar to the example cited. Greater accuracy than that afforded by the curves can be secured by direct substitution in the equations for $\breve{s}$ and $\bar{s}$ or by referring to the original tables.

[12] We will recall with interest how closely the observed average, $\bar{s}=.798\sigma$, of the 1000 values of s corresponding to the 1000 samples of four herein presented checked the theoretical average of $.801\sigma$.

The Alkali Metal Photoelectric Cell

By HERBERT E. IVES

Introduction

IN the development of the commercial system of picture transmission now in operation over certain of the Bell System lines, one of the initial problems was the choice of a method of transforming the light and shade of the picture to be transmitted into properties of an electric current. There are in general two methods of accomplishing this. The first, which we may term the photo-mechanical method, utilizes some photographic process to produce a mechanical structure, which may be used either to make and break contact, or to produce mechanical movement of some element whose motion produces a variable electric current. The second method consists in the utilization of some light sensitive device which produces or varies an electric current.

An indispensable requirement in the electrical transmission of pictures is *speed* in conveying the picture from one point to another. The choice of a method of transforming light and shade into an electrical current will therefore, other things being equal, be that method which requires the least time for the transformation. It is on this basis that the photo-mechanical methods were not favorably considered in this development. The preparation of the line or dot structure image, similar to the half tone plate, or the preparation of a photo-relief, are processes which cannot be completed in less than one to two hours, and involve a delay which in many cases would seriously detract from the advantages of electrical transmission over other means now available, such as the airplane.

In choosing a photo-sensitive device for this purpose, certain requirements had to be met. The light responsive device should be as nearly as possible instantaneous in its action. The response should also be proportional to the light intensity. These requirements cannot be met by any photo-sensitive devices of the group whose resistance changes under the action of light, such as selenium. The field was therefore limited to the *photoelectric cell*, of the type in which the effect of light is to release electrons from the surface of the light sensitive element and so cause an electric current to flow in the space between the light sensitive surface and another electrode. Photoelectric cells are considerably less sensitive than the best variable photo-resistances, but while this characteristic would have made them difficult to utilize in the earlier days of efforts at picture transmission, the development of vacuum tube amplifiers admirably fitted for amplifying photo-

electric currents has remedied this deficiency. A further requirement, that the light sensitive device should preferably be sensitive to visible radiation, ruled out the use of those sensitive materials sensitive chiefly to infra-red or ultra-violet radiation. All of these requirements pointed to the *alkali metal photoelectric cell* as developed by the work of Elster and Geitel and others.

General Characteristics of Photoelectric Cells

The typical photoelectric cell consists of a hermetically sealed glass bulb containing an atmosphere of gas at a low pressure, and provided with two electrodes, one of which is the light sensitive material. In the schematic cell shown in Figure 1, K is the photo-sensitive material

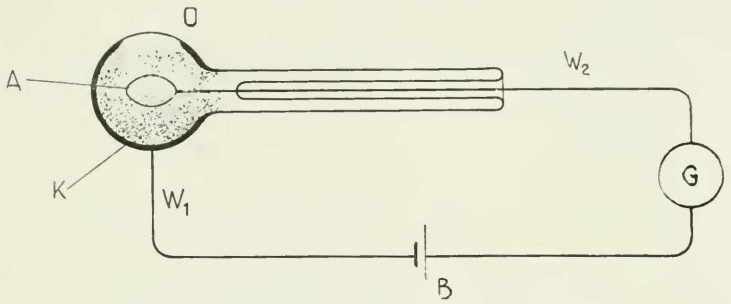


Fig. 1—Schematic central anode photoelectric cell.

(cathode), for instance an alkali metal such as potassium, which is spread upon the inside wall of the glass bulb and is connected with the exterior of the bulb by a sealed-in wire, w_1, A is the other electrode (anode). As here shown, it is a simple metallic ring connected with a second wire, w_2, carried through the stem of the bulb. The two electrodes are shown connected together through a battery, B, and galvanometer, G. The operation of the cell consists in letting light fall upon the cathode through the window, O. The resulting current may then be measured by the galvanometer, or utilized to operate suitable apparatus.

A complete study of the photoelectric cell resolves itself into obtaining knowledge of the effect of varying a number of factors which enter into its construction and use. Of these we may note: the material which is used for the light sensitive surface, and the treatment to which this material is subjected; the composition and pressure of the gaseous atmosphere; the shape and disposition of the various elements, that is, the *structure* of the cell. We must investigate the relationship

between the quality of the light falling upon the cell and the electric current produced. We must in addition consider certain other physical variables which must be met with in practice, notably temperature.

Characteristics of Cells of Various Structures

For purposes of discussion we may classify photoelectric cells in regard to structure as *central cathode cells* and *central anode cells*. As the terms imply, these two extreme types of cells differ in the position of the photoelectric material. The central anode type of cell is shown in Figure 1. The sensitive material entirely covers the walls, so that the cathode is of relatively large area. The central cathode cell is illustrated in Figure 2, in which the symbols are the same as in Figure 1.

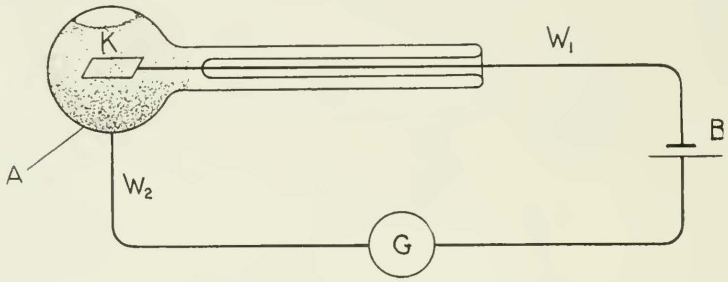


Fig. 2—Schematic central cathode photoelectric cell.

In this the walls of the cell are covered with a non-light sensitive material (e.g. silver), and the sensitive material is coated upon a relatively small centrally placed electrode.

Central Cathode Cells

Central cathode cells possess certain decided advantages for the theoretical study of photoelectric phenomena and have consequently been used in many of the more important photoelectric investigations. The simplest case to consider first is that of the high vacuum cell, that is one containing no appreciable amount of gaseous atmosphere. When a constant light is incident on the light sensitive cathode, and a series of voltages are applied to the terminals of the cell, voltage-current relationships are obtained of the character shown by any one of the curves of Figure 3. Several significant points are to be noted about these characteristic curves. We find that the photoelectric current starts at a definite positive value of the voltage. This voltage is called the "stopping potential." It varies with the wave length of the

exciting light. This is shown in the figure by the several curves for different wave lengths, varying from λ_1, representing short wave energy, such as blue light, to λ_3, long wave energy such as yellow. The shorter the wave length (the higher the frequency), of the exciting light, the higher must be the positive potential necessary to prevent or stop the emission of electrodes under illumination. As the positive potential is reduced, the photoelectric current increases, until the applied field (or the effective field if contact potential differences are present) becomes zero. At this point, the current becomes *saturated*, that is increase of voltage in the negative direction fails to increase the current. This means that the applied field does not penetrate to any appreciable depth into the photoelectric material.

Characteristic curves of the type shown in Figure 3, have played a

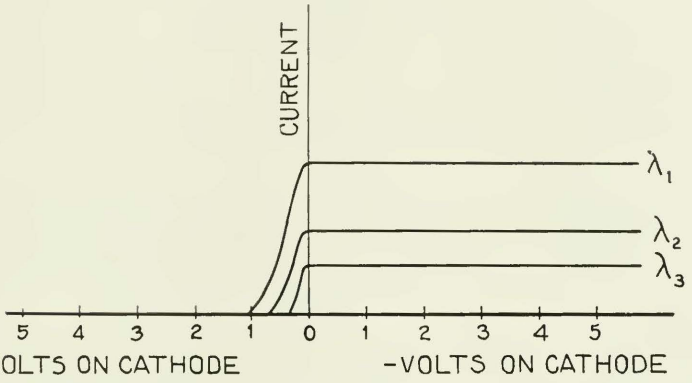


Fig. 3—Voltage-current curves for typical central cathode vacuum photoelectric cell.

very important part in the development of photoelectric theory, and particularly of the quantum theory. If V is the voltage applied to the cell, e the charge on an electron, h the quantum constant, ν the frequency of the exciting light, Einstein predicted and Millikan has shown experimentally that the following relationship holds: $eV = h(\nu - \nu_0)$, where ν_0 is the limiting frequency corresponding to the long wave length limit, beyond which the photoelectric emission does not occur. If m is the mass of the electron, and v its velocity, the above relation can be written $1/2\ m\ v^2\ (\text{Velocity}^2) = h(\nu - \nu_0)$. From this expression it is evident that the greater the interval between the frequency of the light used, and the limiting frequency, the higher is the velocity of emission of the photoelectrons.

When instead of being highly exhausted, the cell has an atmosphere of gas at a low pressure (a few tenths of a millimeter of mercury) the

condition of saturation typical of the high vacuum cell for high negative voltages no longer holds. Instead the photoelectric current is increased by the occurrence of ionization, by collision of the electrons initially produced with the molecules of gas. The current increases with applied voltage in the manner shown in Figure 4, until at some value characteristic of the kind of gas in the cells, the gas breaks down and a visible electrical discharge takes place. The amplifying effect of the gaseous atmosphere increases with the pressure of the gas up

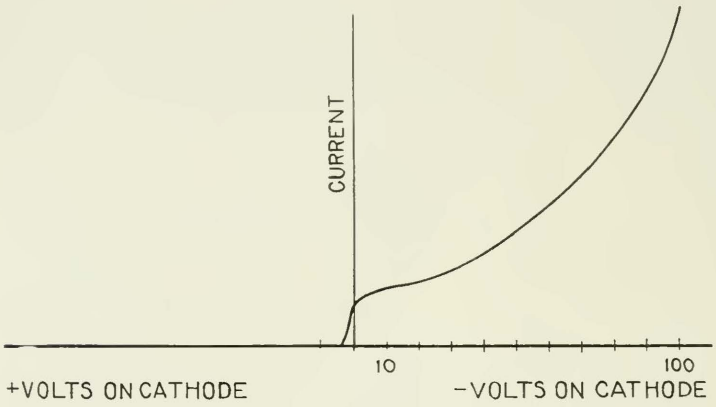


Fig. 4—Voltage-current curves for typical central cathode gas filled photoelectric cell.

to a maximum and then decreases. The value of this optimum pressure depends on the kind of gas and the dimensions of the tube. The best pressure is usually a few tenths of a millimeter of mercury.

As the illumination of the cell is changed, the current changes in exact proportion, that is the illumination-current relationship is rectilinear. This relationship holds for both the vacuum and gas cells provided there are no free glass surfaces on which charges may accumulate. If the window is made too large it may become charged and cause an appreciable curvature of the illumination-current relationship.

Central Anode Cells

In cells with a relatively small centrally placed anode, the voltage-current relationship differs from that of the central cathode cells most noticeably in that high applied voltages are necessary in order to insure saturation. Typical voltage current curves for short (λ_1) and long (λ_2) wave length energy, for a central anode cell consisting of a

spherical anode and concentric spherical cathode are shown in Figure 5. The rate at which saturation is approached with voltage varies with the wave length of the exciting light. The longer the wave length, the slower the electrons, as shown above, and the more quickly are they captured by the central anode.

When gas is introduced into a central anode cell, we again have ionization by collision, and the voltage-current curve is turned upward

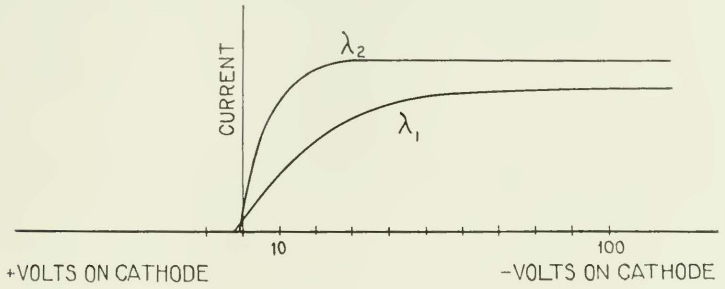


Fig. 5—Voltage-current curves for typical central anode vacuum cell.

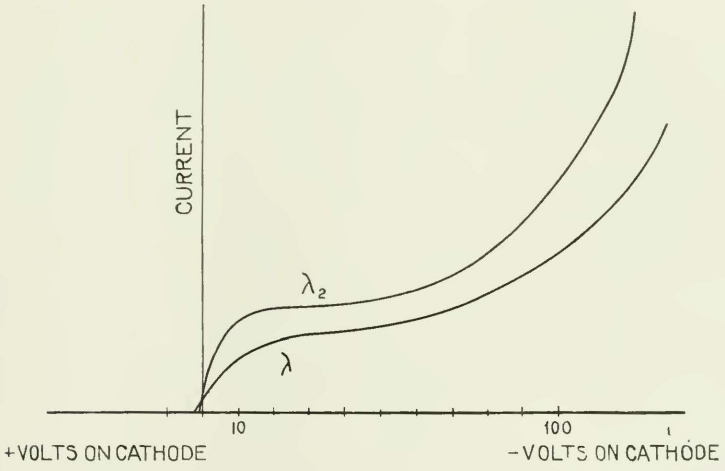


Fig. 6—Voltage-current curves for typical central anode gas filled cell.

from the voltage axis in the manner shown in Figure 6. As in the case of the central cathode cells, the current increases with voltage until the critical potential for the gas is reached.

The illumination-current relationship is rectilinear in the central anode cells, as it was in the central cathode cells, provided the precautions as to avoiding free glass surfaces, already mentioned, are observed.

Influence of the Nature of the Irradiation

The magnitude of the photoelectric current depends upon the angle of incidence, the plane of polarization and the color or wave length of the light used. This dependence is closely interlinked with the choice of the photo-sensitive material and the state of its surface. For the purpose of separating out the effects of the several variables in the incident light, it is necessary to study the properties of optically plane or specular surfaces of the photo-sensitive material. Such surfaces can only be obtained with the alkali metals by raising them above their melting points, by forming alloys, or by depositing extremely thin films on a polished underlying metal surface such as platinum. Specular surfaces of the aklali metals obtained in these several different ways exhibit differences in their behavior, but it will be sufficient for the present purposes to disregard these secondary differences and to speak merely of the photoelectric current from specular surfaces when the incident light varies in wave length in certain typical ways, or is polarized.

Influence of the Plane of Polarization

When light is incident at a steep angle on a specular surface, the two extreme conditions of polarization are those in which the electric vector lies in the plane of incidence and that in which it lies perpendicular to the plane of incidence. In the first case, the electric vector has a component perpendicular to the surface. In the latter case the electric vector lies parallel to the surface. It has long been known that the amount of light absorbed by a metal surface is, in general, greater when the electric vector is in the plane of incidence. Consequently, since photoelectric emission must be due primarily to the absorption of the energy from the incident light, it is to be expected that the photoelectric current will be greater for light polarized with the electric vector in the plane of polarization. Such is actually the case, but while the ratio of absorption of light, for the two planes of polarization, at say 60° incidence, never rises above a value of four for any of the alkali metals, the ratio of the photoelectric currents under the same conditions may mount to a very high value, such as 20 or 30 to one. This effect is particularly noticeable in the liquid alloy of sodium and potassium, and in the case of all the four alkali metals, sodium, potassium, rubidium and caesium, when these spontaneously deposit in a high vacuum upon a polished surface. It is very much less marked in the case of the pure alkali metals in the molten condition. Typical examples of the influence of the plane of polarization

on the photoelectric effect are shown in Figure 7, where the symbol ‖ indicates that the electric vector is in the plane of incidence, the symbol ⊥ that it is perpendicular to this plane.

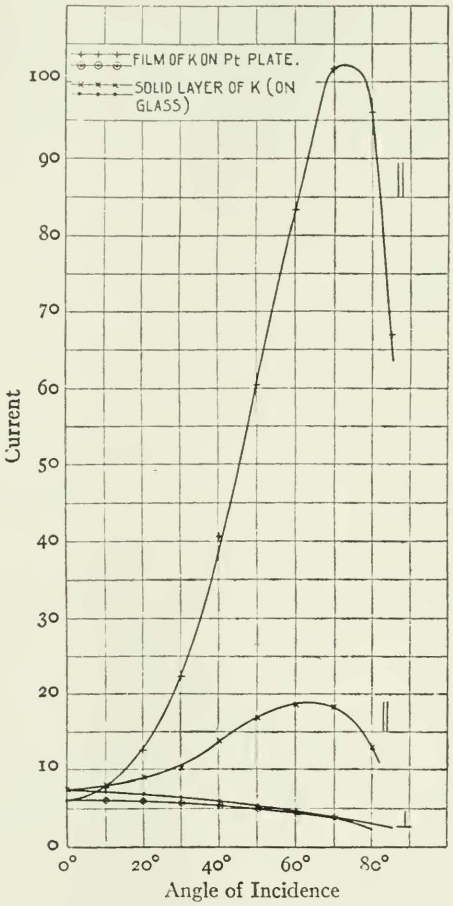


Fig. 7—Photoelectric emission from a specular surface of solid potassium, and from a thinly coated platinum plate, at various angles of incidence; electric vector in plane of incidence (‖); electric vector perpendicular to the plane of incidence (⊥).

Distribution of Response According to Wave Length

When the exciting light is incident either perpendicularly on a specular alkali metal surface, or at a high angle of incidence with the plane of polarization such that the electric vector is parallel to the surface, the response for equal intensities of monochromatic radiation

through the spectrum is as shown by the curve marked ⊥ in Figure 8. Photoelectric emission is entirely absent on the long wave side of a certain wave length which is known as the *long wave length limit.* From this wave length, the emission rises gradually and uniformly toward the short wave or blue end of the spectrum.

When the incident light is polarized so that the electric vector has a component perpendicular to the surface, the wave length distribution of response follows the general character shown in the curve marked ‖ in Figure 8. Correlating the wave length distribution of response with

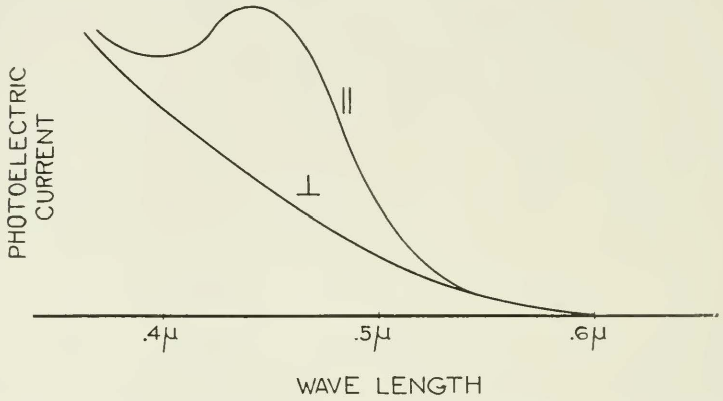


Fig. 8—Wave-length distribution of response from specular alkali metal surface; electric vector in plane of incidence (‖); electric vector perpendicular to plane of incidence (⊥).

the variation of emission with the plane of polarization considered in the last section, the general conclusion may be drawn that the enhanced emission for light with the electric vector in the plane of incidence is due largely to radiation falling within a narrow spectral region. The magnitude of this wave length peak varies greatly with different materials. The maximum occurs at a wave length different not only for the different alkali metals, but also for different modes of securing the specular surface. This wave length maximum has the appearance of being due to some resonance phenomenon, and its variation in position through the spectrum according to the method of preparation of the surface is connected in some unknown way with the state of binding of the alkali metal atom on the surface with the body of material beneath.

The Photoelectric Current from Rough Surfaces

With rough surfaces of alkali metal, the plane of polarization of the incident light no longer has meaning. We would therefore expect no

significant difference in the total emission from a really rough surface when the plane of polarization of the incident light is changed; and such in fact is the case. We would also expect that the distribution of response according to wave length would be a mixture of the effects of the two planes of polarization. This also is found to be so. Rough surfaces show the wave length maximum characteristic of light polar-

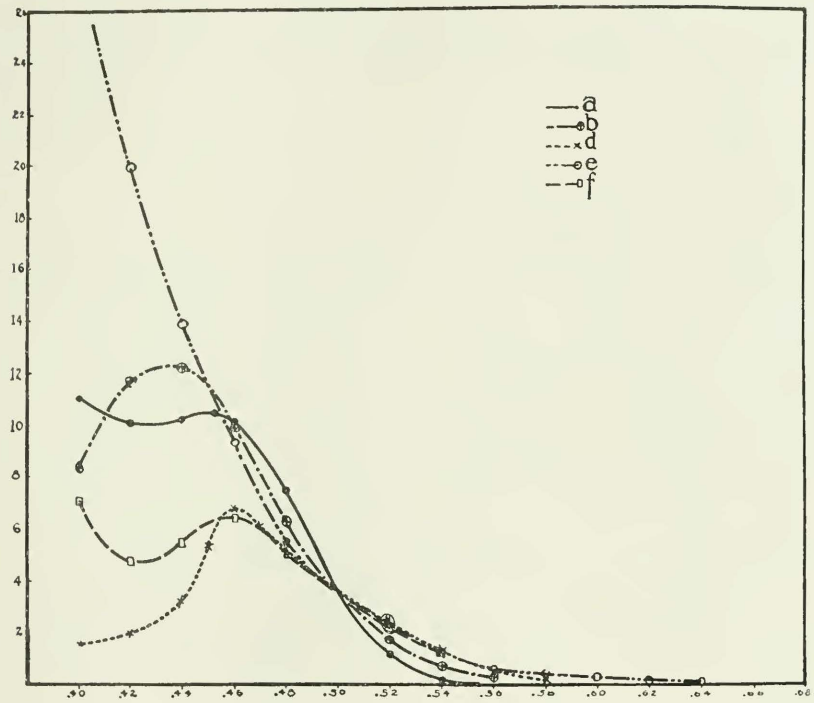


Fig. 9—Wave-length distribution of emission from rough surface potassium photoelectric cells, showing variation from cell to cell depending on difference of treatment.

ized with the electric vector in the plane of incidence on a specular surface. Depending on the degree of roughness and the method of preparation of the surface, the wave length maximum is different in size and also in position. In Figure 9 are shown several wave length distribution curves for potassium cells in which the surface is roughened and colored by a hydrogen glow discharge. These cells differ both in respect to their absolute sensitiveness, in the position of their maximum sensitiveness, and the extent of their sensitiveness toward the red end of the spectrum. It has not as yet been found possible to prepare photoelectric cells with properties uniform from one cell to another.

Influence of the Photoelectric Material and its Treatment

The alkali metals differ in their photoelectric sensitiveness in a perfectly definite order, which is that of their degree of electro-positiveness, as shown by their position in the periodic table of the elements. The variation in sensitiveness is correlated with the extension of sensitiveness in the spectrum. This progresses regularly from sodium, which in its pure state is not photoelectrically sensitive beyond about .58μ, through potassium and rubidium, to caesium, which is photoelectrically sensitive in the near infra-red. The exact terminations of sensitiveness in the spectrum depend upon the character of the surface and its treatment, and have not been exactly correlated with any other properties of the material.

In order to attain the greatest sensitiveness with the alkali metals, these are commonly subjected, in the preparation of the photoelectric cell, to what is called the *coloring* process, discovered by Elster and Geitel. This consists in subjecting the surface to a glow discharge in an atmosphere of hydrogen. The result is to color the otherwise silvery alkali metal a rather deep blue-purple or blue-green. The exact cause of this color is not known, but it has every appearance of being due to the production of small (colloidal) particles of alkali metal. The greater sensitiveness is probably due to the increased effective surface presented by the colloidal particles rather than the increased absorption coefficient of the darker color. Similar colors may be obtained by distilling the alkali metal in a very thin layer on glass and the color of the surface changes when observed by polarized light in much the same manner as do colloidal surfaces of other sorts.

After the completion of the coloring process, it is necessary to remove all the hydrogen from the cell by pumping. Otherwise the surface will revert to its original uncolored form. In order to obtain the amplifying effect of a gaseous atmosphere, it is customary to introduce an inert gas, such as argon or helium, into the cell.

Cells made in the manner just outlined are reasonably permanent in their important characteristics. Elster and Geitel have made potassium cells in this manner, which when connected with a delicate electrometer exhibited a degree of sensitiveness approximately that of the human eye. According to what has gone before, the most sensitive cells should be obtained if rubidium or caesium are used in place of potassium. It is found however by experiment that rubidium, and particularly caesium, do not lend themselves so well to the coloring

process, probably because of their lower melting points, and hence cells made of potassium may represent practically the best performance which is now attainable.

Effect of Temperature

It has long been held that the photoelectric effect is independent of temperature. Recent experiments, however, have shown that the alkali metals are affected in their photoelectric response by variation of temperature. A typical set of data for potassium is shown in Figure 10. It will be noted that the influence of temperature is small

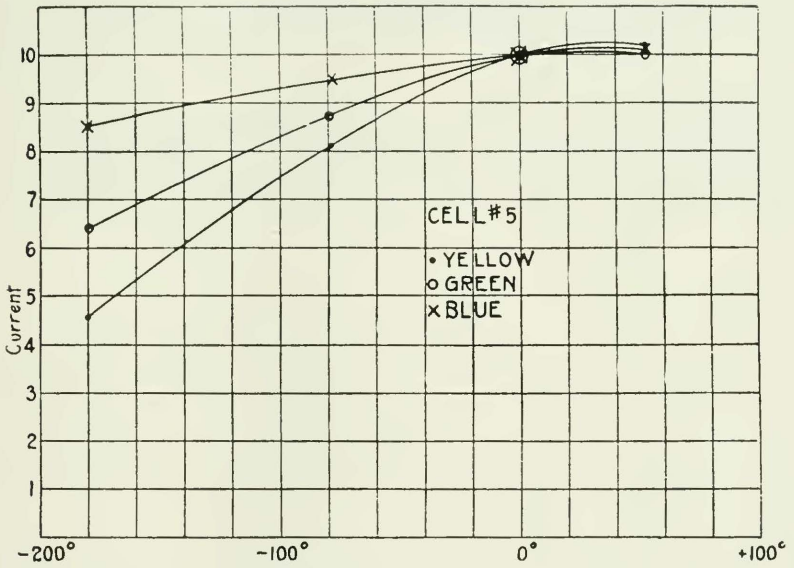


Fig. 10—Variation of photoelectric sensitiveness of potassium with temperature.

for the shorter wave lengths of light, but considerable for long wave excitation. These data were secured in highly exhausted cells of pure alkali metal. When gas is present or a large amount of alkali metal vapor can deposit on the cooled surface, the effect of decreased temperature may be to increase the photoelectric current. It will be noted that the changes shown in Figure 10 are insignificant over the ordinates range of room temperatures, and for practical purposes, particularly for picture transmission, the effects of temperature on the performance of photoelectric cells may be taken as negligible.

Practical Features of Cells as Used

The photoelectric cells as used for picture transmission are classified, according to the above discussion, as central anode, gas filled, colored cells. The shape of the cells is that shown in Figure 1, and also in the photograph, Figure 11, with which is an accompanying scale. The cells are made of pyrex glass, which is chosen because it is highly resistant to corrosion by potassium during the distillation stages. The very long neck of the cells is dictated partly by the space into which the cells are placed in the picture transmission apparatus, in

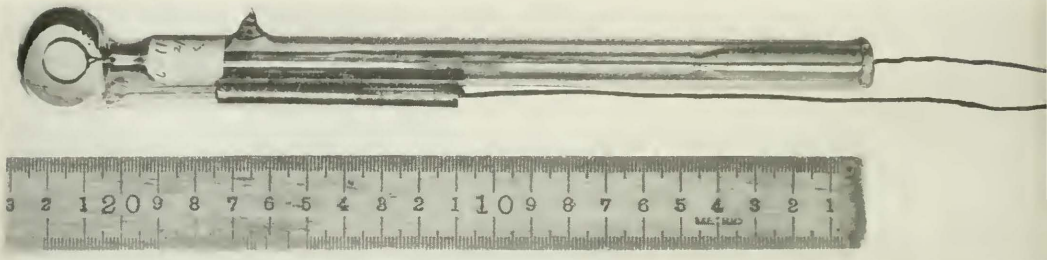


Fig. 11—Photograph of photoelectric cell of type used in picture transmission.

part by the desirability of having as long an insulating space as possible. Where, as in earlier types of photoelectric cells, the alkali metal is in close proximity to the other electrode, leakage currents over the glass surface greatly interfere with accurate results. (In working with extremely small currents it is desirable to have in addition to the considerable length of glass insulating path, a metallic guard ring in the stem of the cell, which may be earthed.)

The alkali metal ordinarily used is potassium. This is introduced by distillation on the pump. The cell is first baked to a temperature of 400° C. for several hours while on the pump in order to drive out all traces of water vapor. The potassium for use in making up the photoelectric cells is first of all distilled in a vacuum into long glass tubes. In this preliminary distillation, the greater part of the absorbed gaseous impurities are removed. After the cell has been baked out on the pump, a piece of the glass tube containing potassium is broken off and introduced into the pump system. Between the point of introduction and the cell are a series of bulbs. The potassium after melting in vacuo is distilled successively through these bulbs and into the photoelectric cell, where it is condensed on the walls of the bulb. A window is then made in the cell by applying a small flame on the appropriate

part. The next step is to introduce a small amount of pure hydrogen gas, which is permitted to enter from a reservoir on the system. This hydrogen gas goes through the system of bulbs through which the potassium has been distilled, which still contain a large amount of potassium, and is thereby cleaned of all traces of gases or vapors which might react on the potassium in the cell. A glow discharge is then

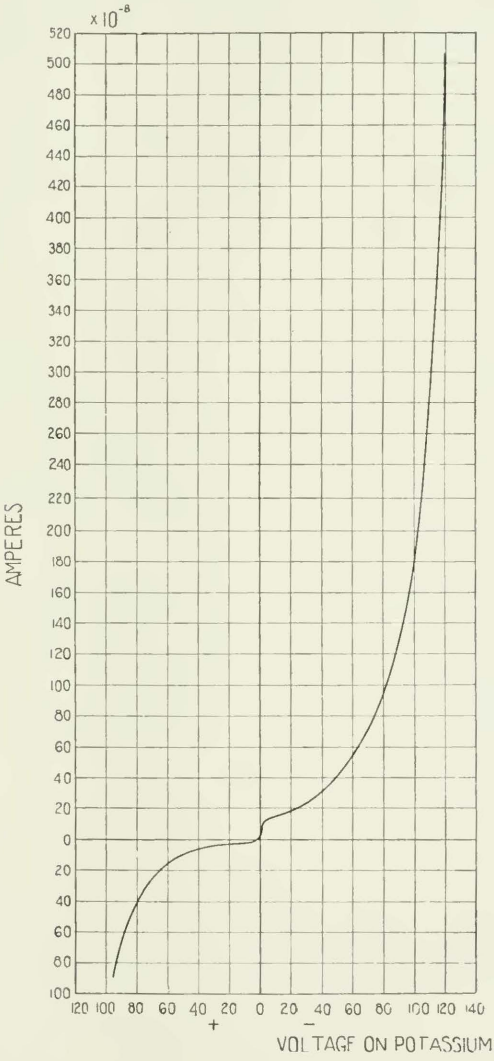


Fig. 12—Typical voltage current characteristic of potassium photoelectric cell as used in picture transmission. Incident luminous flux = .015 lumen.

passed from a high voltage source, until, by illuminating the alkali metal surface and reading the current on a sensitive galvanometer, it is found that a maximum of sensitiveness has been attained. The hydrogen is then completely removed by long continued pumping. The final step in the preparation of the cell consists in the introduction of a small quantity of carefully purified argon. The argon for this purpose is held in a reservoir in which there is a pool of sodium-potassium alloy. By passing an electric discharge from this pool to an electrode through the gas, the argon is purified of all active impurities. It is introduced into the cell through the same series of potassium coated bulbs already mentioned, the potassium in the meantime having been vigorously heated to drive off all occluded hydrogen, so that the gas when it finally reaches the photoelectric cell is entirely inert. The gas pressure is carefully adjusted while the cell is still on the pump so as to give an optimum effect, after which the cell is sealed off.

Typical voltage-current characteristics of the cells thus made are as shown on Figure 12, where the currents indicated are those obtained from an illumination of 100 meter candles from a vacuum tungsten lamp, the aperture of the cell being 1.5 sq. cm. It will be noted that, unlike the ideal characteristic shown in Figure 6, the actual cells shows a small current in the opposite direction for positive voltages applied to the sensitive surface. This is because practically it is very difficult to prevent some alkali metal from depositing on the anode, which thus becomes light sensitive, and responds to the scattered and reflected light in the cell.

For use in the picture transmission apparatus, the cells are mounted in tubular metal cases, from which they are insulated by hard rubber rings attached to the glass stem, by sealing wax. The cells in their cases are handled as units; and are sufficiently rugged to be readily shipped from place to place. Their characteristics remain practically unchanged indefinitely.

SELECTED BIBLIOGRAPHY

Books

Photo-electricity, H. Stanley Allen, 2nd Ed. 1925 (containing bibliography of articles up to 1925).

Photo-electricity, A. L. Hughes, 1914.

Die Lichtelektrische Erscheinungen, Pohl and Pringsheim, 1914.

National Research Council Report on Photo-electricity, A. L. Hughes, 1921.

Recent Articles

The Transmission of Pictures over Telephone Lines, H. E. Ives, J. W. Horton, R. D. Parker and A. B. Clark, Bell System Technical Journal, April, 1925.

Photo-electric Properties of Thin Films of Alkali Metals, H. E. Ives, Astrophysical Journal *60*, p. 209, 1924.

The Normal and Selective Photo-electric Effects in the Alkali Metals and their Alloys, H. E. Ives and A. L. Johnsrud, Astrophysical Journal *60*, p. 231, 1924.

Variation with Temperature of the Photo-electric Effect in Potassium Photo-electric Cells, H. E. Ives, Optical Society of America Journal *8*, p. 551, 1924.

The Influence of Temperature on the Photo-electric Effect of the Alkali Metals, H. E. Ives and A. L. Johnsrud, Optical Society of America Journal *11*, p. 565, 1925.

The Voltage-Current Relation in Central Anode Photo-electric Cells, Ives and Fry, Astrophysical Journal *66*, p. 1, 1922.

Electric Circuit Theory and the Operational Calculus[1]

By JOHN R. CARSON

CHAPTER IX

The Finite Line with Terminal Impedances

So far in our discussions of wave propagation in lines and wave-filters, we have confined attention to the case where the impressed voltage is applied directly to the infinitely long line. We have found that, by virtue of this restriction, the indicial admittance functions of the important types of transmission systems are rather easily derived and expressible in terms of well known functions, and the essential phenomena of wave propagation clearly exhibited. In practice, however, we are concerned with lines of finite length with the voltage impressed on the line through a terminal impedance Z_1 and the distant end closed by a second terminal impedance Z_2. We now take up the problem presented by such a system.

Let $K=K(p)$ denote the characteristic operational impedance of the line, and $\gamma=\gamma(p)$ the operational propagation constant of the line. We have then

$$\begin{aligned} V &= Ae^{-\gamma x}+Be^{\gamma x}, \\ I &= \frac{1}{K}Ae^{-\gamma x}-\frac{1}{K}Be^{\gamma x}, \end{aligned} \tag{240}$$

where A and B are so far arbitrary constants. To determine these constants we assume an e.m.f. E impressed on the line at $x=0$ through a terminal impedance Z_1 and the line closed at $x=s$ by a second terminal impedance Z_2. At $x=s$ we have therefore

$$Z_2 I = V$$

whence from (240)

$$\frac{Z_2}{K}e^{-\gamma s}A-\frac{Z_2}{K}e^{\gamma s}B=Ae^{-\gamma s}+Be^{\gamma s}$$

and

$$B=-\frac{1-\rho_2}{1+\rho_2}e^{-2\gamma s}A \tag{241}$$

where $\rho_2=Z_2/K$.

[1] Concluded from the issue of January, 1926.

At $x=0$ we have

$$V=E-Z_1I$$

whence

$$A+B=E-\frac{Z_1}{K}A+\frac{Z_1}{K}B,$$

$$(1+\rho_1)A+(1-\rho_1)B=E, \tag{242}$$

where $\rho_1=Z_1/K_1$.

From (241) and (242) we get

$$A=\frac{1+\rho_2}{(1+\rho_1)(1+\rho_2)-(1-\rho_1)(1-\rho_2)e^{-2\gamma s}}E$$

$$B=\frac{-(1-\rho_2)e^{-2\gamma s}}{(1+\rho_1)(1+\rho_2)-(1-\rho_1)(1-\rho_2)e^{-2\gamma s}}E$$

and finally

$$I_x=\frac{E}{K+Z_1}\frac{e^{-\gamma x}+\dfrac{1-\rho_2}{1+\rho_2}e^{-\gamma(2s-x)}}{1-\dfrac{1-\rho_1}{1+\rho_1}\dfrac{1-\rho_2}{1+\rho_2}e^{-2\gamma s}}. \tag{243}$$

If we replace E by a unit e.m.f. we get the operational formula for the indicial admittance A_x; thus

$$A_x=\frac{\lambda}{K}\frac{e^{-\gamma x}+\mu_2e^{-\gamma(2s-x)}}{1-\mu_1\mu_2e^{-2\gamma s}}=\frac{1}{Z_x(p)} \tag{244}$$

where

$$\lambda=K/(K+Z_1),$$

$$\mu_1=\frac{1-\rho_1}{1+\rho_1}=\frac{K-Z_1}{K+Z_1},$$

$$\mu_2=\frac{1-\rho_2}{1+\rho_2}=\frac{K-Z_2}{K+Z_2}.$$

K,γ,Z_1,Z_2,μ_1 and μ_2 are, of course, functions of the operator p.

The integral equation corresponding to the operational formula (244) is

$$\frac{1}{pZ_x(p)}=\int_0^\infty e^{-pt}A_x(t)dt. \tag{245}$$

Now by (244) we can expand $1/Z_x(p)$; it is

$$\begin{aligned}\frac{1}{Z_x(p)} &= \lambda\frac{e^{-\gamma x}}{K} + \lambda\mu_2\frac{e^{-\gamma(2s-x)}}{K} \\ &+ \lambda\mu_1\mu_2\frac{e^{-\gamma(2s+x)}}{K} + \lambda\mu_1\mu_2^2\frac{e^{-\gamma(4s-x)}}{K} \\ &+ \lambda\mu_1^2\mu_2^2\frac{e^{-\gamma(4s+x)}}{K} + \cdots\end{aligned} \tag{246}$$

Now we observe that $e^{-\gamma x}/K$ is simply the operational formula for the indicial admittance at point x of an infinitely long line with unit e.m.f. impressed directly on the line at $x=0$. This will be denoted by $a_x(t)$. Similarly $e^{-\gamma(2s-x)}/K$ is the operational formula for the indicial admittance at point $(2s-x)$ with unit e.m.f. impressed directly on the line at $x=0$. This will be denoted by $a_{2s-x}(t)$, etc.

Recognition of this fact allows us to derive a formal solution in terms of a series of reflected waves. For let a set of functions $v_o, v_1, v_2, v_3, \ldots$ satisfy and be defined by the operational equations

$$\begin{aligned} v_o &= \lambda(p) = \lambda \\ v_1 &= \lambda\mu_2 \\ v_2 &= \lambda\mu_1\mu_2 \\ v_3 &= \lambda\mu_1\mu_2^2, \text{ etc.}\end{aligned} \tag{247}$$

It then follows from the preceding and theorem II that

$$A_x(t) = \frac{d}{dt}\int_0^t d\tau \left\{ \begin{aligned} &v_o(t-\tau)a_x(\tau) + v_1(t-\tau)a_{2s-x}(\tau) \\ &+ v_2(t-\tau)a_{2s+x}(\tau) + \cdots \end{aligned} \right\}. \tag{248}$$

If, therefore, we know the indicial admittance of the infinitely long line with unit e.m.f. directly applied and if we can solve the operational equations (247), then $A_x(t)$ is given by (248) by integration. This solution may well present formidable difficulty in the way of computation. It is, however, formally straightforward and the numerical computation is entirely possible, the only question being as to whether the importance of the problem justifies the necessary expenditure of time and effort. Without any computations, however, the solution (248) admits of considerable instructive interpretation by inspection. The first term represents the current at point x of an infinitely long line in response to a unit e.m.f. impressed at $x=0$ through an impedance Z_1; $v_o = v_o(t)$ is the corresponding voltage across the line terminals proper. The second term is a reflected wave from the other

terminal due to the terminal irregularity which exists there. The third term is a reflected wave from the sending end terminal, etc. The solution is therefore a wave solution and is expanded in a form which corresponds exactly with the sequence of phenomena, which it represents.

The solution takes a particularly instructive form when $Z_1 = k_1 K$ and $Z_2 = k_2 K$ where k_1 and k_2 are numerics. Then

$$\begin{aligned} v_0 &= \frac{1}{1+k_1} \\ v_1 &= \frac{1}{1+k_1}\,\frac{1-k_2}{1+k_2} \\ v_2 &= \frac{1}{1+k_1}\,\frac{1-k_2}{1+k_2}\,\frac{1-k_1}{1+k_1}, \text{ etc.} \end{aligned} \tag{249}$$

and

$$A_x(t) = \frac{1}{1+k_1}\left\{ \begin{array}{l} a_x(t) + \dfrac{1-k_2}{1+k_2}\, a_{2s-x}(t) \\ \quad + \dfrac{1-k_1}{1+k_1}\,\dfrac{1-k_2}{1+k_2}\, a_{2s+x}(t) + \ldots . \end{array} \right\}. \tag{250}$$

If $k_1 = 0$, $k_2 = 1$ we have the case of the e.m.f. impressed directly on the sending end of the line and the distant end closed through its characteristic impedance; the solution reduces to

$$A_x(t) = a_x(t)$$

as, of course, it should be by definition.

If $k_1 = 0$ and $k_2 = \infty$, we have the case of the line open-circuited at the distant end, and the solution reduces to

$$A_x(t) = \tfrac{1}{2}\{a_x(t) - a_{2s-x}(t) - a_{2s+x}(t) + a_{4s-x}(t) + \ldots\}. \tag{251}$$

Finally, if both k_1 and k_2 are zero, the line is shorted and

$$A_x(t) = \tfrac{1}{2}\{a_x(t) + a_{2s-x}(t) + a_{2s+x}(t) + a_{4s-x}(t) + \ldots\}. \tag{252}$$

The operational equations (247) admit of further interesting and instructive physical interpretation without computation. Consider a circuit consisting of an impedance Z_1 in series with an impedance K. Let a unit e.m.f. be applied to this circuit and let v_0 be the re-

sultant voltage across the impedance K. Then, operationally,

$$v_o = \frac{K}{K+Z_1} = \lambda$$

so that v_o, thus defined in physical terms, is the v_o of equations (247).

Now let this voltage be impressed on a circuit consisting of an impedance $2Z_2$ in series with an impedance $K-Z_2$ so that the total impedance is $K+Z_2$. Let the resultant voltage drop across the impedance element $K-Z_2$ be denoted by v_1; then operationally

$$v_1 = \frac{K}{K+Z_1} \cdot \frac{K-Z_2}{K+Z_2} = \lambda\mu_2$$

which agrees with v_1 as given by equation (247).

Similarly if voltage v_1 is applied to a circuit consisting of an impedance $2Z_1$ in series with an impedance $K-Z_1$ and if v_2 denote the voltage drop across impedance $K-Z_1$, then

$$v_2 = \lambda\mu_1\mu_2$$

We can thus see physically what the voltages $v_o, v_1, v_2 \ldots$ mean in terms of simple circuits consisting of K and Z_1 in series and K and Z_2 in series respectively.

I shall now work out a specific problem exemplifying the preceding theory. The example is made as simple as possible for two reasons. First because its simplicity makes it more instructive than when the phenomena depicted and the essentials of the mathematical methods are obscured by complicated formulas and extensive computations. Secondly while the general method of solution illustrated is thoroughly practical we cannot hope to arrive at the numerical solutions of the complicated problems without a large amount of laborious computations. Problems involving transmission lines with complicated terminal impedances are among the most difficult, as regards actual numerical solution, of any which present themselves in mathematical physics. On the other hand, the formal solution (248) gives at a glance the essential character of the phenomena involved.

The specific problem we shall deal with may be stated as follows: A unit e.m.f. is directly impressed on the terminals of a transmission line of length s, the distant end of which is closed by a condenser C_o. The line is supposed to be non-dissipative, its constants being inductance L and capacity C per unit length. Required the current at any point $x (x<s)$ of the line.

We write $\sqrt{L/C}=k$, $1/\sqrt{LC}=v$: then by virtue of the preceding

analysis of transmission line propagation the indicial admittance a_x of the *infinitely long line* is given by

$$a_x = 0, \text{ for } t < x/v,$$

$$= \frac{1}{k}, \text{ for } t \geqq x/v.$$

The operational characteristic impedance is, of course, $k = \sqrt{L/C}$, and the terminal impedances Z_1 and Z_2 are given by

$$Z_1 = 0,$$

$$Z_2 = 1/pC_o.$$

Referring now to equation (244) we have:—

$$\lambda = 1, \quad \mu_1 = 1,$$

$$\mu_2 = \frac{k - 1/pC_o}{k + 1/pC_o} = \frac{kC_o p - 1}{kC_o p + 1}.$$

Consequently, referring to equations (247), we have, operationally,

$$v_o = 1$$

$$v_1 = v_2 = \frac{kC_o p - 1}{kC_o p + 1}$$

$$v_3 = v_4 = \left(\frac{kC_o p - 1}{kC_o p + 1}\right)^2$$

$$v_5 = v_6 = \left(\frac{kC_o p - 1}{kC_o p + 1}\right)^3.$$

In order to determine these functions we have therefore to solve the general operational equation

$$V_n = \left(\frac{kC_o p - 1}{kC_o p + 1}\right)^n$$

where V_n denotes either v_{2n-1} or v_{2n}.

In order to eliminate the coefficient kC_o, we make use of theorem VIII, and write

$$\phi_n = \left(\frac{p-1}{p+1}\right)^n.$$

In accordance with that theorem

$$V_n(t) = \phi_n(t/kC_o).$$

We therefore start with the operational equation

$$\phi_n = \left(\frac{p-1}{p+1}\right)^n.$$

Now the solution of this operational equation is very easy and can be expressed in a number of ways. We require it expressed in the form most easily computed. The following appears best adapted for our purposes. Consider the auxiliary operational equation:—

$$\sigma_n = \left(\frac{p-2}{p}\right)^n$$

$$= \left(p^n - 2\frac{n}{1!}p^{n-1} + 2^2\frac{(n)(n-1)}{2!}p^{n-2}\right.$$

$$\left. + \ldots + (-1)^n 2^n\right)\frac{1}{p^n}.$$

The explicit solution is gotten by replacing $1/p^n$ by $t^n/n!$ and p^n by d^n/dt^n, whence

$$\sigma_n(t) = \left(\frac{d^n}{dt^n} - 2\frac{n}{1!}\frac{d^{n-1}}{dt^{n-1}} + \ldots + (-1)^n 2^n\right)\frac{t^n}{n!},$$

$$= 1 - \frac{n}{1!}\frac{2t}{1!} + \frac{n(n-1)}{2!}\frac{(2t)^2}{2!} + \ldots + (-1)^n\frac{(2t)^n}{n!}.$$

But writing

$$\sigma_n = \left(\frac{p-2}{p}\right)^n = \frac{1}{H(p)}$$

it follows that

$$\phi_n = \left(\frac{p-1}{p+1}\right)^n = \frac{1}{H(p+1)}$$

$$= \frac{p+1}{p}\cdot\frac{p}{p+1}\frac{1}{H(p+1)}$$

$$= \left(1+\frac{1}{p}\right)\cdot\frac{p}{p+1}\frac{1}{H(p+1)}.$$

Referring now to theorem VII, we see that

$$\phi_n(t) = \left(1 + \int_0^t dt\right)\sigma_n(t)\cdot e^{-t}.$$

Since we have already solved for $\sigma_n(t)$, this determines $\phi_n(t)$ and hence $V_n(t)$. The functions $v_0, v_1, v_2 \ldots$ are therefore determined.

Now refer back to equation (248) giving the required current in terms of $v_o, v_1, v_2 \ldots$ and the admittances $a_x(t), a_{2s-x}(t), \ldots$. It follows at once by substitution of the preceding that

$$A_x(t) = \frac{1}{k}\left\{ v_o\left(t-\frac{x}{v}\right)+v_1\left(t-\frac{2s-x}{v}\right)+v_2\left(t-\frac{2s+x}{v}\right)+\ldots\right\}$$

the functions v_o, v_1, v_2 being zero for negative values of the argument. This result may possibly require a little explanation.

Consider the expression

$$\frac{d}{dt}\int_0^t f(t-\tau)\cdot 1(\tau)d\tau$$

where $1(t)$ denotes a function which is zero for $t<t_o$ and unity for $t\geq t_o$. It is evidently identical with the admittance $a_x(t)$ provided the proper value is assigned to t_o.

Now since $1(t)=0$ for $t<t_o$ and unity for $t\geq t_o$, the preceding may be written as zero for $t<t_o$, and

$$\frac{1}{k}\frac{d}{dt}\int_{t_o}^t f(t-\tau)d\tau \qquad \text{for } t\geq t_o$$

which is equal to $f(t-t_o)$.

If we set $x=0$, we get the current entering the line; thus

$$\begin{aligned} A_o(t) &= \frac{1}{k}\left\{ v_o(t)+v_1\left(t-\frac{2s}{v}\right)+v_2\left(t-\frac{2s}{v}\right)\right. \\ &\qquad \left. +v_3\left(t-\frac{4s}{v}\right)+v_4\left(t-\frac{4s}{v}\right)+\ldots\right\} \\ &= \frac{1}{k}\left\{ 1+2V_1\left(t-\frac{2s}{v}\right)+2V_3\left(t-\frac{4s}{v}\right)\right. \\ &\qquad \left. +2V_5\left(t-\frac{6s}{v}\right)+\ldots.\right\}. \end{aligned}$$

This has been computed for the case where $\sqrt{sLC_o}=10$ and is shown in Fig. 26. Referring to this figure we see that the current jumps at $t=0$ to the value $\sqrt{C/L}=1/k$, and keeps this constant value for a time interval $2s/v$. At this instant the first reflected wave arrives and the current takes another jump, of $2/k$. Thereafter it begins to decrease very slowly until time $t=4s/v$ at which time it takes another

jump of $2/k$. Thereafter we have a series of jumps of $2/k$ at time intervals $2s/v$, the current decreasing between successive jumps. The smooth curve is the indicial admittance of an oscillation circuit consisting of an inductance sL in series with a capacity C_o. We see therefore, that the current in the line oscillates with discontinuous

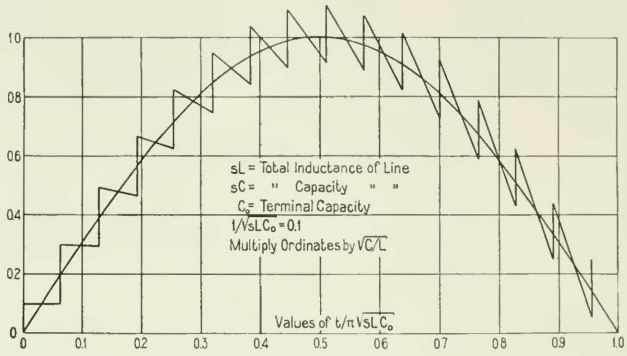


Fig. 26—Current entering non-dissipative line terminated by capacity C_o unit E.M.F. applied to line

jumps about the current in the corresponding oscillation circuit. Since the whole circuit contains no resistance, the oscillations never die away, but continue to oscillate, as shown, about the curve

$$\sqrt{\frac{C}{L}} \sin\left(\frac{t}{\sqrt{sLC_o}}\right)$$

which is the indicial admittance of the corresponding oscillation circuit.

I shall now discuss a method of solving circuit theory problems, quite generally applicable to complicated networks, and particularly useful in dealing with transmission lines terminated in impedances. I have found it particularly useful in arriving at numerical solutions where other methods prove far more laborious. It is also of mathematical interest, as it applies another type of integral equation to the problems of electric circuit theory.

Suppose that we have a network with two sets of terminals as shown in Fig. 27.[7] Now suppose that terminals 22 are short circuited and a unit e.m.f. inserted between terminals 11. Let the resultant current flowing between terminals 11 be denoted by $S_{11}(t) = S_{11}$ and that

[7] Regarding conventions as to signs, see the Appendix to this chapter.

between terminals 22 by $S_{21}(t) = S_{21}$. S_{11} is the driving point indicial admittance with respect to terminals 11 and S_{21} the transfer indicial admittance of terminals 22 with respect to 11 under short circuit conditions.

Similarly if terminals 11 are shortcircuited and a unit e.m.f. inserted



Fig. 27

between terminals 22 the current flowing between terminals 22 is denoted by $S_{22}(t) = S_{22}$ and that flowing between terminals 11 by $S_{12}(t) = S_{12}$. If the network is passive, i.e., contains no internal source of energy, it follows from the reciprocal theorem that $S_{21} = S_{12}$. As far as the two sets of terminals are concerned, the network is completely specified by the indicial admittances $S_{11}, S_{22}, S_{21} = S_{12}$.

Now let a voltage $V_1(t) = V_1$ be inserted between terminals 11, and a voltage $V_2(t) = V_2$ between terminals 22. The current flowing between terminals 11, denoted by I_1 is

$$I_1(t) = \frac{d}{dt}\int_0^t V_1(\tau) S_{11}(t-\tau)\,d\tau + \frac{d}{dt}\int_0^t V_2(\tau) S_{12}(t-\tau)\,d\tau \qquad (253)$$

while the corresponding current between terminals 22 is

$$I_2(t) = \frac{d}{dt}\int_0^t V_2(\tau) S_{22}(t-\tau)\,d\tau + \frac{d}{dt}\int_0^t V_1(\tau) S_{21}(t-\tau)\,d\tau \qquad (254)$$

Now consider two networks of indicial admittances S_{11}, S_{22}, $S_{12} = S_{21}$ and $T_{11}, T_{22}, T_{12} = T_{21}$ respectively and let them be connected in tandem as shown in Fig. 28 to form a compound network.

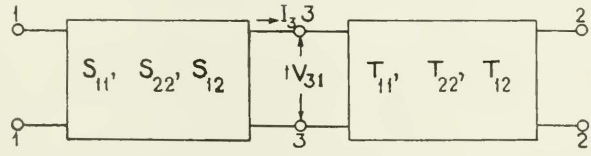


Fig. 28

We require the indicial admittances of the compound network in terms of the indicial admittances of the component networks.

Short circuit terminals 22 of the compound network and insert a

unit e.m.f. between terminals 11. Let $V_{31}(t)$ denote the resultant voltage between terminals 33 measured in the direction of the arrow, and I_3 the current flowing between the networks. We have then the two following expressions for the current I_3.

$$I_3 = S_{21}(t) - \frac{d}{dt}\int_0^t V_{31}(\tau) S_{22}(t-\tau) d\tau \tag{255}$$

and

$$I_3 = \frac{d}{dt}\int_0^t V_{31}(\tau) T_{11}(t-\tau) d\tau. \tag{256}$$

Equating we get

$$\frac{d}{dt}\int^t V_{31}(\tau) \left[S_{22}(t-\tau) + T_{11}(t-\tau)\right] d\tau = S_{21}(t). \tag{257}$$

By precisely similar reasoning, if terminals 11 are short circuited and a unit e.m.f. inserted between terminals 22, and the corresponding voltage across terminals 33 denoted by V_{32}, we have [8]

$$\frac{d}{dt}\int_0^t V_{32}(\tau) \left[S_{22}(t-\tau) + T_{11}(t-\tau)\right] d\tau = T_{12}(t). \tag{258}$$

Equations (257) and (258) are integral equations of the Poisson type which completely determine V_{31} and V_{32} in terms of the indicial admittances S and T. We shall discuss the solution of these equations presently.

If $U_{11}, U_{22}, U_{21} = U_{12}$ denote the indicial admittances of the compound network we have at once

$$U_{11} = S_{11}(t) - \frac{d}{dt}\int_0^t V_{31}(\tau) S_{12}(t-\tau) d\tau \tag{259}$$

$$U_{22} = T_{22}(t) - \frac{d}{dt}\int_0^t V_{32}(\tau) T_{21}(t-\tau) d\tau \tag{260}$$

and

$$\begin{aligned} U_{21} = U_{12} &= \frac{d}{dt}\int_0^t V_{31}(\tau) T_{21}(t-\tau) d\tau \\ &= \frac{d}{dt}\int_0^t V_{32}(\tau) S_{12}(t-\tau) d\tau. \end{aligned} \tag{261}$$

If, therefore, equations (257) and (258) are solved for V_{31} and V_{32}, the required indicial admittances of the compound network are given

[8] V_{32} being opposite to V_{31} in direction.

by (259), (260) and (261) in terms of the indicial admittances of the component networks.

A simple example will now be worked out illustrating the method of solution just discussed. Suppose that a unit e.m.f. is impressed on a transmission line (infinitely long) of distributed constants R,L,C, through a terminal resistance R_o. Required the terminal line voltage V.

The operational equation of this problem is gotten in the usual manner. The current entering the line is

$$V\sqrt{\frac{Cp}{Lp+R}}.$$

It is also obviously equal to $\frac{1}{R_o}(1-V)$: equating the two expressions, and rearranging we get:—

$$V=\frac{\sqrt{\frac{Lp+R}{Cp}}}{R_o+\sqrt{\frac{Lp+R}{Cp}}}.$$

Writing $R/2L=\rho$ and setting $R_o=\sqrt{L/C}$, this becomes

$$V=\frac{\sqrt{1+2\rho/p}}{1+\sqrt{1+2\rho/p}}. \tag{262}$$

This operational equation can, of course, be solved in a number of ways, though, as a matter of fact, its numerical solution is quite troublesome. This point will be returned to later: we shall first formulate the problem in accordance with the method just discussed.

The indicial admittance of the line is known; it is

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

Consequently the current entering the line is explicitly

$$\frac{d}{dt}\int_0^t V(\tau)A(t-\tau)d\tau.$$

But the current is also equal to $\frac{1}{R_o}(1-V(t))$; equating, we get

$$V(t)=1-R_o\frac{d}{dt}\int_0^t V(\tau)A(t-\tau)d\tau.$$

Performing the indicated differentiations

$$V(t) = 1 - R_o A(o)\, V(t) - R_o \int_0^t V(\tau) A'(t-\tau) d\tau.$$

Now $A(o) = \sqrt{\frac{C}{L}}$ and

$$A'(t) = \rho e^{-\rho t}(I_1(\rho t) - I_o(\rho t)) \sqrt{\frac{C}{L}}$$

and $R_o = \sqrt{L/C}$; therefore the equation becomes

$$V(t) = \frac{1}{2} + \frac{\rho}{2} \int_0^t V(t-\tau)[I_o(\rho\tau) - I_1(\rho\tau)] e^{-\rho\tau} d\tau.$$

As a matter of convenience we change the time scale to ρt, and get

$$V(t) = \frac{1}{2} + \frac{1}{2} \int_0^t V(t-\tau)[I_o(\tau) - I_1(\tau)] e^{-\tau} d\tau \tag{263}$$

$$= \frac{1}{2} - \frac{1}{2} \int_0^t d\tau\, V(t-\tau) \frac{d}{d\tau} e^{-\tau} I_o(\tau),$$

where it is understood that t is actually ρt. This is the integral equation of the problem and is in the canonical form of Poisson's integral equation.

Before solving this equation numerically I shall show how a simple approximate solution is obtainable immediately; an advantage often attaching to this type of integral equation.

The function $\frac{d}{dt} e^{-t} I_o(t)$ is equal to -1 for $t = 0$ and converges rapidly to zero. $V(t)$ has, as we know from the operational equation, the initial value 1/2 and the final value 1. Neither function changes sign. It follows from the mean value theorem that the equation can be written as

$$V(t) = \frac{1}{2} - \frac{1}{2} V(t) \int_0^{\alpha t} \frac{d}{dt} e^{-t} I_o(t) dt$$

where $\alpha \leq 1$. Integrating

$$V(t) = \frac{1}{2} - \frac{1}{2} V(t)\, [e^{-\alpha t} I_o(\alpha t) - 1]$$

and

$$V(t) = \frac{1}{1 + e^{-\alpha t} I_o(\alpha t)}. \tag{264}$$

The correct initial and final values of $V(t)$ result for all final values of $\alpha \leq 1$; so that approximately

$$V(t) = \frac{1}{1+e^{-t}I_o(t)}.$$

This equation, while not exact, except for $t=0$ and t very large, shows faithfully the general character of $V(t)$ and the way it approaches its final value unity. For large values of t

$$e^{-t}I_o(t) = 1/\sqrt{2\pi t}$$

whence

$$V(t) = \frac{1}{1+1/\sqrt{2\pi t}}, \qquad t \geq 8. \tag{264-a}$$

Approximations of the foregoing type are not always possible and may not be of sufficient accuracy. I shall therefore give next a method of numerical solution which is generally applicable to integral equations of this type and works quite well in practice. We shall write the integral equation in the more general form

$$u(x) = f(x) + \int_0^x u(x-y)k(y)dy \tag{265}$$

where $f(x)$ and $k(y)$ are known and $u(x)$ unknown. The method depends on the numerical integration of the definite integral. Let us divide the x scale into small intervals d and for convenience write

$$u(nd) = u_n$$
$$f(nd) = f_n$$
$$k(nd) = k_n.$$

Now from the integral equation we have at once

$$u(o) = u_o = f_o,$$
$$u(d) = u_1 = f_1 + \int_0^d u(d-y)k(y)dy.$$

Now if d is taken sufficiently small

$$\int_0^d u(d-y)k(y)dy = \frac{d}{2}[u_1k_o + u_ok_1],$$

whence

$$u_1 = f_1 + \frac{d}{2}[u_1k_o + u_ok_1]$$

and

$$u_1 = \frac{1}{1-k_od/2}[f_1 + u_ok_1d/2]$$

which determines u_1 since u_o is known. Similarly

$$u_2 = f_2 + d\left[\tfrac{1}{2}u_o k_2 + u_1 k_1 + \tfrac{1}{2}u_2 k_o\right]$$

which determines u_2. Proceeding in the same manner

$$u_3 = f_3 + d\left[\tfrac{1}{2}u_o k_3 + u_1 k_2 + u_2 k_1 + \tfrac{1}{2}u_3 k_o\right], \text{ etc.}$$

In this way we determine the value of $u(x)$, point by point from the recurrence formula

$$u_n = \frac{f_n + d\left[\tfrac{1}{2}u_o k_n + u_1 k_{n-1} + u_2 k_{n-2} + \cdots + u_{n-1} k_1\right]}{1 - \tfrac{1}{2}k_o d}. \tag{266}$$

The result of the application of numerical integration, in accordance with formula (266), to the integral equation (263) is shown in Fig. (29). The dotted curve is a plot of the approximate solution as

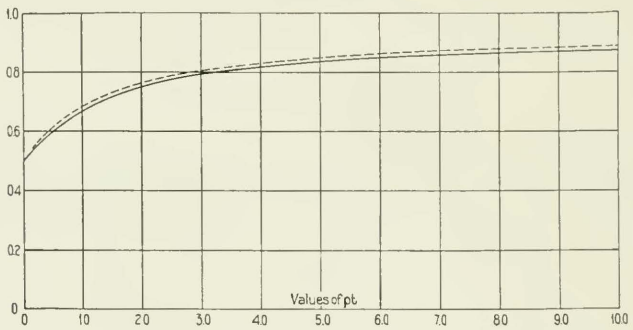


Fig. 29—Line terminal voltage unit E.M.F. impressed on line through resistance $R_0 = \sqrt{L/C}$

given by equation (264), for $\alpha = 1$. We see that the voltage starts with the value 1/2 and slowly reaches its ultimate value, unity, its approach to unity, for large values of t, being in accordance with the formula

$$V(t) = \frac{1}{1 + 1/\sqrt{2\pi t}}.$$

The application of the foregoing method to the transmission line problem proceeds as follows. Let $S_{11}(t)$, $S_{22}(t)$ and $S_{12}(t)$ be the short indicial admittances of the line. $S_{11}(t)$ is the current entering the line (at $x = 0$) with unit e.m.f. directly impressed and the distant end short circuited. $S_{12}(t)$ is the current at $x = s$ under the same

circumstances. Consequently from (252)

$$\begin{aligned} S_{11}(t) &= a_o(t) + 2a_{2s}(t) + 2a_{4s}(t) + \ldots \\ S_{12}(t) &= 2\{a_s(t) + a_{3s}(t) + a_{5s}(t) + \ldots\}. \end{aligned} \tag{267}$$

S_{22} is clearly equal to S_{11} by symmetry.

Now suppose that an e.m.f. $E = E(t)$ is impressed on the line at $x = 0$, $t = 0$, through a terminal impedance Z_1, and the distant end $(x = s)$ closed through an impedance Z_2. We suppose these terminal impedances and the actual impressed e.m.f. replaced by the actual line voltages V_1 and V_2, impressed directly on the line at $x = 0$ and at $x = s$ are

$$\begin{aligned} I_o(t) = &\frac{d}{dt}\int_0^t S_{11}(t-\tau)\, V_1(\tau)\, d\tau \\ &- \frac{d}{dt}\int_0^t S_{12}(t-\tau)\, V_2(\tau)\, d\tau, \end{aligned} \tag{268}$$

$$\begin{aligned} I_s(t) = &-\frac{d}{dt}\int_0^t S_{12}(t-\tau)\, V_1(\tau)\, d\tau \\ &+ \frac{d}{dt}\int_0^t S_{22}(t-\tau)\, V_2(\tau)\, d\tau. \end{aligned} \tag{269}$$

But the current at $x = s$ is also equal to the current in the terminal impedance Z_2 in response to the terminal voltage V_2: denoting by $\alpha_2(t)$ the indicial admittance of Z_2 it is

$$I_s(t) = \frac{d}{dt}\int_0^t \alpha_2(t-\tau)\, V_2(\tau)\, d\tau. \tag{270}$$

Similarly the current entering the line at $x = 0$ is the current flowing in the terminal impedance Z_1 in response to the e.m.f. $E - V_1$. Denoting by $\alpha_1(t)$ the indicial admittance of Z_1, it is

$$I_o(t) = \frac{d}{dt}\int_0^t \alpha_1(t-\tau)\{E(\tau) - V_1(\tau)\}\, d\tau. \tag{271}$$

Equating equation (268) and (271) and (269) and (270) we eliminate $I_o(t)$ and $I_s(t)$ and get

$$\begin{aligned} &\int_0^t [S_{11}(t-\tau) + \alpha_1(t-\tau)]\, V_1(\tau)\, d\tau - \int_0^t S_{12}(t-\tau)\, V_2(\tau)\, d\tau \\ &\qquad = \int_0^t \alpha_1(t-\tau) E(\tau)\, d\tau, \end{aligned} \tag{272}$$

$$-\int_0^t S_{12}(t-\tau)\, V_1(\tau)\, d\tau + \int_0^t [S_{22}(t-\tau) - \alpha_2(t-\tau)]\, V_2(\tau)\, d\tau = 0. \tag{273}$$

These two equations are simultaneous integral equations of the Poisson type in V_1 and V_2, which completely determine these voltages provided the admittances and the impressed voltages are known. They therefore represent the application of a new type of integral equation to the problem of electric circuit theory.

The numercial solution of the general case, either by (248) or (272–273) is necessarily laborious when the terminal impedances are complicated and is only justified when the technical importance of the problem is considerable. I wish, however, to emphasize two points in this connection: the numerical solution is always entirely possible and, compared with other and older forms of solution, enormously simpler. One has only to inspect the classical forms of solution of problems of the type to realize the truth of this last statement.

I shall now give two applications of equations (272–273) to specific problems, in one of which an approximate solution of the integral equation can be gotten, and in the other of which numerical integration is applied.

Problem I. Given a non-inductive cable of distributed constants C and R and length s, with unit e.m.f. applied at $x=0$, while at $x=s$ the cable is closed by a condenser C_o. Required the terminal voltage $V(t)$ across the condenser C_o.

We first write down the short-circuit indicial admittances of the cable; from equation (168) of a preceding section and equation (267) they are:—

$$S_{11}(t)=S_{22}(t)$$

$$=\sqrt{\frac{C}{\pi Rt}}\left\{1+2e^{-\frac{4\beta}{t}}+2e^{-\frac{16\beta}{t}}+2e^{-\frac{36\beta}{t}}+\ldots\right\}, \tag{274}$$

$$S_{12}(t)=S_{21}(t)$$

$$=2\sqrt{\frac{C}{\pi Rt}}\left\{e^{-\frac{\beta}{t}}+e^{-\frac{9\beta}{t}}+e^{-\frac{25\beta}{t}}+\ldots\right\}, \tag{275}$$

where $\beta=s^2RC/4$.

Now the current at $x=s$ is equal to

$$S_{12}(t)-\frac{d}{dt}\int_0^t V(\tau)S_{22}(t-\tau)d\tau.$$

It is also the condenser current due to the voltage $V(t)$; that is

$$C_o\frac{d}{dt}V(t).$$

Equating the two expressions and integrating we get

$$C_o V(t) = \int_0^t S_{12}(\tau)d\tau - \int_0^t V(\tau)S_{22}(t-\tau)d\tau \tag{276}$$

which is the integral equation of the problem. In order to get an approximate solution without detailed computation we assume that the cable is long. In this case the leading terms of (274) and (275) are large compared with the terms following: Furthermore $S_{12}(t)$ builds up very slowly while $S_{22}(t)$ is a rapidly varying function. A good approximation therefore results if we take $V(\tau)$ outside the integral sign in (276) and write

$$C_o V(t) = \int_0^t S_{12}(\tau)d\tau - V(t)\int_0^t S_{22}(\tau)d\tau$$

whence

$$V(t) = \frac{1}{C_o}\frac{\int_0^t S_{12}(t)dt}{1+\frac{1}{C_o}\int_0^t S_{22}(t)dt}. \tag{277}$$

This approximation is quite good for long cables and shows the way $V(t)$ builds up quite truthfully. We see that V is initially zero, and builds up ultimately to unity. For large values of t, it becomes

$$V(t) = \frac{\int_0^t S_{12}(t)dt}{\int_0^t S_{22}(t)dt}. \tag{278}$$

This is the approximate formula also for the open circuit voltage, as may be seen by setting $C_o = 0$ in (277).

In electric circuit problems, it is often sufficient, as implied above, to know qualitatively the behavior of an electric system without going through the labor of detailed computation. For this purpose the formulation of the problem as a Poisson Integral Equation is particularly well adapted. A simple example will be given, which can be checked from the known solution. Suppose that we require the voltage V at point x of an infinitely long transmission line (L,R,C) in response to a unit e.m.f. impressed at $x=0$. This is, of course, known from formula (211-a): we shall here be concerned, however, with approximate solutions from the Poisson integral equation of the problem.

If $a_x(t)$ denote the indicial admittance of the line at point x, then the current at point x is simply $a_x(t)$, which is given by formula (210-a). But if $V(t)$ is the voltage at point x, the current is also given by

$$\frac{d}{dt}\int_0^t V(\tau)a_o(t-\tau)d\tau.$$

Equating these two expansions, we get the integral equation of the problem

$$\frac{d}{dt}\int_0^t V(\tau)a_o(t-\tau)d\tau = a_x(t).$$

Now if we write $T=\rho t-A$ where $\rho=R/2L$ and $A=\frac{xR}{2}\sqrt{\frac{C}{L}}$, then

$$a_x=\sqrt{\frac{C}{L}}e^{-(T+A)}I_o\sqrt{T(T+2A)}, \quad T\geq 0,$$

and in terms of the relative time T, the integral equation is reducible to

$$\frac{d}{dT}\int_0^T V(T-\tau)e^{-\tau}I_o(\tau)d\tau = e^{-(T+A)}I_o(\sqrt{T(T+2A)}$$

while the exact formula for V is by (211-a)

$$V(T)=e^{-A}+Ae^{-A}\int_0^T e^{-\tau}I_1\left(\sqrt{\tau(\tau+2A)}\right)\frac{}{\sqrt{\tau(\tau+2A)}}d\tau.$$

From the integral equation it is easy to establish superior and inferior limits for $V(T)$; it is

$$V(T)\leq e^{-A}\frac{I_o(\sqrt{T(T+2A)}}{I_o(T)} = V_a(T),$$

$$\geq \frac{\int_0^T e^{-T}\, I_o(T)\, V_a(T)dT}{\int_0^T e^{-T}\, I_o(T)dT} = V_b(T).$$

Both formulas give the correct initial and final values of V; namely e^{-A} and unity. Since V lies between V_a and V_b, the mean value $(V_a+V_b)/2$ also has correct initial and final values and should be a better approximation than either. The table given below shows the orders of approximation obtainable from the case where $A=3$. It is evident from this table that the foregoing approximate formulas exhibit the form of $V(T)$ qualitatively in a quite satisfactory manner.

T	V_a	V_b	$\frac{1}{2}(V_a+V_b)$	V
0	0.05	0.05	0.05	0.05
2	0.25	0.12	0.18	0.17
4	0.39	0.19	0.29	0.26
6	0.50	0.23	0.36	0.32
8	0.57	0.27	0.42	0.37
10	0.64	0.31	0.47	0.41
12	0.69	0.34	0.51	0.44
15	0.74	0.38	0.56	0.48
18	0.78	0.41	0.60	0.52

Problem II. Our second illustrative problem may be stated as follows:—A unit e.m.f. is impressed on a transmission line of length s and distributed constants L,R,C. At $x=s$ the line is closed by a resistance R_o in parallel with an inductance L_o. Required the current in the terminal resistance. If $V(t)$ denotes the terminal voltage, the current at $x=s$ is given by

$$S_{12}(t) - \frac{d}{dt}\int_0^t V(\tau)S_{22}(t-\tau)d\tau.$$

It is also equal to the current flowing into the terminal impedance; that is

$$\frac{1}{R_o}V(t) + \frac{1}{L_o}\int_0^t V(\tau)d\tau.$$

Equating and rearranging

$$\left[\frac{1}{R_o}+S_{22}(o)\right]V(t) = S_{12}(t) - \int_0^t V(\tau)\left[\frac{1}{L_o}+S'_{22}(t-\tau)\right]d\tau. \quad (279)$$

Now the short circuit admittance S_{22} and S_{12} are given by formula (210-a) of a preceding chapter, and $S_{22}(o)=\sqrt{C/L}$. In order to apply numerical integration to (279), numerical values must be assigned to the constants. We take

$$R_o=\sqrt{L/C}=1935 \text{ ohms},$$

$$L_o=0.4 \text{ henry},$$

$$\frac{R}{2L}=\rho=292,$$

$$v=1/\sqrt{LC}=1.105\times10^4,$$

$$s=100.$$

The results of the numerical evaluation of equation (279), with these values inserted, is shown in Fig. 30. The voltage is identically zero until $vt=100$; $t=100/v$ is the time of propagation of the line. At that instant it jumps to the value $e^{-\rho t}=e^{-100\rho/v}$ and then begins to die away rapidly due to the draining action of the inductance.

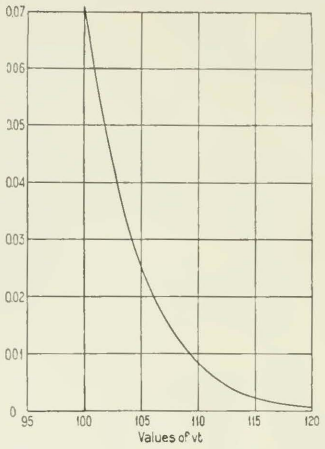


Fig. 30—Voltage across terminal impedance on smooth line

The effect of secondary reflection is insignificant and therefore not shown. The current in the terminal resistance is V/R_o so that it is given by the same curve.

I have reserved until the last the exposition of the *expansion theorem solution* as applied to transmission lines with terminal impedances, for the reason that it is the least powerful and the most restricted, although most closely resembling the classical form of solution. Furthermore, it does not represent the sequence of physical phenomena, in fact it is not a *wave* solution, but a solution in terms of normal or characteristic vibration. In practical application its usefulness is restricted to the non-inductive cable.

It will be recalled that the expansion theorem solution is formulated as follows:—

If

$$A=1/Z(p)$$

is the operational equation of the problem, then the explicit solution is

$$A(t)=\frac{1}{Z(o)}+\sum_{1}^{n}\frac{e^{p_k t}}{p_k Z'(p_k)}$$

where $p_1, p_2 \ldots$ are the roots of the equation $Z(p)=0$.

Let us apply this formula to the case of a line of length s, with unit e.m.f. directly applied at $x=s$, and line short circuited at $x=s$. Referring to equation (244) and putting $\lambda=\mu_1=\mu_2=1$ we get

$$A_x=\frac{1}{K}\frac{\cosh\gamma(s-x)}{\sinh\gamma s}=\frac{1}{Z_x(p)} \tag{280}$$

as the operational formula of the problem. This can be written as

$$A_x=(Cp+G)\frac{\cosh\gamma(s-x)}{\gamma\sinh\gamma s}=\frac{1}{Z_x(p)} \tag{281}$$

where in the general case,

$$\gamma=\sqrt{(Lp+R)(Cp+G)}. \tag{282}$$

The values of γ for which $Z_x(p)$ vanishes are the roots of the transcendental equation

$$\sinh\gamma s=0$$

excluding zero. These roots are infinite in number: Let γ_m be the m^{th} root; then

$$\gamma_m=i\frac{m\pi}{s},\quad m=1,2,\ \ldots\ \infty. \tag{283}$$

The corresponding values of p_m are then gotten by solving (282) for p and writing $\gamma=\gamma_m$.

The explicit solution of the operational equation (281) is then

$$\begin{aligned}A_x(t)&=\frac{1}{Z_x(o)}+\sum\left(C+\frac{G}{p_m}\right)\frac{\cosh\gamma_m(s-x)}{s\gamma_m\dfrac{d\gamma_m}{dp_m}\cosh\gamma_m s}e^{p_m t}\\&=\frac{1}{Z_x(o)}+\sum(C+G/p_m)\frac{\cosh\gamma_m x}{s\gamma_m\dfrac{d\gamma_m}{dp_m}}e^{p_m t}.\end{aligned} \tag{284}$$

Let us apply this to the non-inductive, non-leaky cable in which $L=G=0$ and $\gamma=\sqrt{RCp}$, so that

$$p_m=\gamma_m^2/RC=-\frac{m^2\pi^2}{s^2RC},$$

and

$$\gamma_m\frac{d\gamma_m}{dp_m}=\frac{RC}{2},$$

Also $Z_x(o) = sR$. We thus get

$$A_x(t) = \frac{1}{sR} + \frac{2}{sR} \sum_{m=1}^{\infty} \cos \frac{m\pi}{s} x . e^{-\frac{m^2\pi^2}{s^2RC} t} . \tag{285}$$

This is a thoroughly practical formula for computation, owing to the rapid convergence of the series. In fact, for this particular line termination chosen, it is probably the simplest and most easily computed form of solution. These advantages depend, however, strictly on two facts. First, the fact that the line is taken as non-inductive and secondly that the terminations chosen are those of a short circuit. In fact, as we shall see, it is only in the case of the non-inductive cable that this type of solution is of any practical value.

There is one other point which should be carefully observed in connection with this solution (285). This is that it is not expressed in terms of a series of direct and reflected waves, corresponding to the sequence of physical phenomena, but in terms of *normal* or *characteristic vibrations*. This point will be returned to later.

Let us now attempt to apply this type of solution to the transmission line, L,R,C,G. Writing

$$\rho = \frac{R}{2L} + \frac{G}{2C},$$

$$\sigma = \frac{R}{2L} - \frac{G}{2C},$$

$$v = 1/\sqrt{LC}.$$

We have

$$\gamma^2 = \frac{1}{v^2}[(p+\rho)^2 - \sigma^2]$$

whence

$$p_m = -\rho \pm v \sqrt{\gamma_m^2 + \frac{\sigma^2}{v^2}}$$

$$= -\rho \pm iv \sqrt{\left(\frac{m\pi}{s}\right)^2 - \frac{\sigma^2}{v^2}}, \qquad m = 1,2, \ldots$$

$$\gamma_m \frac{d\gamma_m}{dp_m} = \frac{1}{v^2}(p_m + \rho)$$

$$= \pm \frac{i}{v} \sqrt{\left(\frac{m\pi}{s}\right)^2 - \frac{\sigma^2}{v^2}}.$$

Setting $G=0$ for simplicity and substituting in (284) we get, after easy simplifications,

$$A_x(t)=\frac{1}{sR}+\frac{2vC}{s}\sum\frac{\cos\left(\frac{m\pi}{s}x\right)}{\sqrt{\left(\frac{m\pi}{s}\right)^2-\frac{\rho^2}{v^2}}}\sin\left(vt\sqrt{\left(\frac{m\pi}{s}\right)^2-\frac{\rho^2}{v^2}}\right)e^{-pt}. \quad (286)$$

If we write

$$\sqrt{\left(\frac{m\pi}{s}\right)^2-\frac{\rho^2}{v^2}}=\mu_m\frac{m\pi}{s}$$

(286) can be written as

$$A_x(t)=\frac{1}{sR}+\frac{vC}{s}\sum\frac{e^{-pt}}{\frac{m\pi}{s}\mu_m}\left\{\sin\frac{m\pi}{s}(\mu_m vt-x)+\sin\frac{m\pi}{s}(\mu_m vt+x)\right\}. \quad (287)$$

This type of solution is often referred to as a *wave* solution and the component terms of the series regarded as travelling waves. As a matter of fact it is a solution in terms of normal or characteristic vibrations, each of which is to be regarded as instantaneously produced at time $t=0$. The solution in terms of true waves has been fully discussed in the preceding.

Formula (287) is practically useless for computation on account of the slow convergence of the series (the series are only conditionally convergent), and cannot be interpreted to bring out the existence of the actual direct and reflected waves and the physical character of the phenomena it formulates. In fact, as stated above, this form of solution is useful only in connection with the non-inductive cable.

In the cases considered above we have taken the simplest possible terminations—these of short circuits in which case the roots of $Z(p)$ are easily evaluated. If, however, the line is closed by arbitrary impedances, the case is quite different, and the location of the roots becomes, except for simple impedances, and then only in the case of the non-inductive cable, practically impossible. While, therefore, the expansion theorem solution can be formally written down, its actual numerical evaluation is a practical impossibility, except in a few cases. For this reason it will not be considered further here.

The physically artificial character of the expansion solution, as applied to transmission lines, may be seen from the following considerations. When a wave is sent into the line, for a finite time equal to the time of the propagation of the line, it is independent of the character of the distant termination. Yet in the expansion solution every term involves and is dependent upon the impedance constitut-

ing the distant termination. Evidently, from physical considerations, the series of component vibrations making up the complete solution must therefore so combine as to annihilate the effect of the distant termination for a finite time. The solution is, therefore, mathematically correct but physically artificial.

Note on Integral Equations.

An integral equation is defined as an equation in which the unknown function occurs under a sign of integration; the process of determining the unknown function is called solving the equation.

Integral equations are of great importance in mathematical physics and in recent years very considerable work has been done on them from the standpoint of pure analysis.

The types of integral equations with which we are concerned in the present work are *Laplace's Equation*

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

and *Poisson's Equation*

$$\phi(x) = f(x) + \int_0^x \phi(y)K(x-y)dy.$$

But little work has been done on Laplace's Equation from the standpoint of pure analysis; its most extensive and useful applications appear to be in connection with the Operational Calculus. Practical methods of solution are extensively discussed in the text.

We shall now briefly discuss the solution of Poisson's Equation.

The formal series solution, which is absolutely convergent, is obtained by successive substitution. Thus suppose we write

$$\phi(x) = \phi_0(x) + \phi_1(x) + \phi_2(x) + \ldots .$$

and define the terms of the series in accordance with the scheme

$$\phi_0(x) = f(x),$$

$$\phi_1(x) = \int_0^x \phi_0(y)K(x-y)dy,$$

$$\phi_2(x) = \int_0^x \phi_1(y)K(x-y)dy, \text{ etc.},$$

the resulting series satisfies the integral equation and is absolutely convergent. It is, however, practically useless for computation or interpretation.

A power series solution, when it exists, can be gotten by repeated differentiation; thus

$$\phi(o) = f(o),$$

$$\phi'(x) = f'(x) + \phi(o)K(x) + \int_0^x \phi'(x-y)K(y)dy,$$

$$\phi'(o) = f'(o) + \phi(o)K(o)$$

In this way all the derivatives at $x=0$ are calculable; let them be denoted by $\phi_o, \phi_1, \phi_2 \ldots$ Then

$$\phi(t) = \phi_o + \phi_1 \frac{x}{1!} + \phi_2 \frac{x^2}{2!} + \ldots .$$

This form of solution, also, is of limited practical usefulness, except for small values of x.

A number of mathematicians, including Wittaker and Bateman, have studied the question of numerical solution and suggested other processes. After quite extensive study of the question, however, the writer is of the opinion that point-by-point numerical integration like that discussed in the text is, in general, the most practical, rapid and accurate method of numerical solution. This judgment is confirmed by G. Prasad who, in a paper on the Numerical Solution of Integral Equations delivered before the International Mathematical Congress (Toronto, 1924), discusses the whole question and arrives at the same conclusion.

In the text, numerical integration is carried out in accordance with Simpson's Rule. It is possible, of course, to employ more complicated and refined formulas for approximate quadrature. It is the writer's opinion that this is hardly justified in practical problems and that the required accuracy is more simply obtained by employing smaller intervals.

Appendix to Chapter IX. Note on Conventions as to Signs in Networks

In the network shown on page 196 the arrows indicate the directions chosen as positive in the network itself, quite regardless of the presence of any e.m.fs. and currents.

The sign attributed to a current, an e.m.f., or a voltage is positive if the current, e.m.f., or voltage is in the positive direction; otherwise the sign is negative.

Stated more fully:

A current at a specific point (at a specific instant of time) is posi-

tive if it is flowing in the positive direction; negative if flowing in the negative direction.

An e.m.f. or a voltage between two points is positive if the potential increases in the positive direction between the two points; negative if the potential increases in the negative direction. (It may be noted that this convention makes the sign of a voltage the same as the sign of that e.m.f. which could be inserted between the two points without producing any effects in the network.)

CHAPTER X

Introduction to the Theory of Variable Electric Circuits[9]

In the preceding chapters it has everywhere been assumed that the networks are *invariable*: that is to say, that the constants and connections of the network do not vary or change with time. In many important technical problems, however, we wish to know, not merely what happens when an electromotive force is applied to an invariable network, but the effect of suddenly changing a circuit constant or of introducing a variable circuit element. In the present chapter we shall show that this type of problem can be dealt with by a simple extension of the methods discussed in the preceding chapters.

The simplest and at the same time one of the most technically important problems of this type is the effect of sudden short circuits and sudden open circuits on an energized network or system. This type of problem will serve as an introduction to the more general theory.

The Sudden Short Circuit

Consider the network shown in Fig. 31.

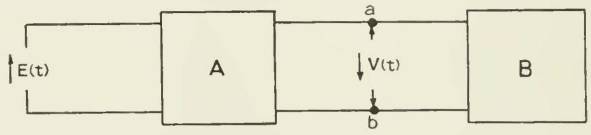


Fig. 31

This network, which for generality is supposed to consist of two parts A and B, indicated schematically, is energized by an electromotive force $E(t)$ which produces a voltage $V(t)$ between the points ab. The voltage $V(t)$ is calculable by usual methods from $E(t)$ and the constants and connections of the network, supposed to be specified.

[9] The material in this chapter is largely taken from a paper by the writer on "Theory and Calculation of Variable Electrical Systems," Phys. Rev. Feb. 1921.

We now suppose that, at reference time $t=0$, a short circuit is suddenly placed across ab; and require the effect of this short circuit on the distributions of currents in the network. The solution of this problem is based on the following proposition:

The effect of the short circuit is precisely the same as the insertion at time $t=0$ of a voltage $-V(t)$, equal and opposite to $V(t)$, between points a and b.

The resultant currents in the system for $t \geq 0$ are then composed of two components:—

(1) The currents which would exist in the invariable network, in the absence of the short circuit, due to the impressed source $E(t)$. These are calculable by usual methods.

(2) The currents due to the electromotive force $V(t)$ inserted at time $t=0$, between the points a and b. These are also calculable by usual methods, since $V(t)$ is itself known from the primary distribution of currents and charges.

By the preceding analysis we have succeeded, therefore, in reducing the problem of a sudden short circuit, to the determination of the currents in an *invariable* network in response to a suddenly impressed electromotive force: that is, the problem to which the preceding chapters have been devoted.

The Sudden Open Circuit

The problem of a sudden open circuit in any part of a network can be dealt with in a precisely analogous manner, although the actual calculation of the resultant current and voltage distribution is mathematically more complicated. Consider the network shown in Fig. 32.

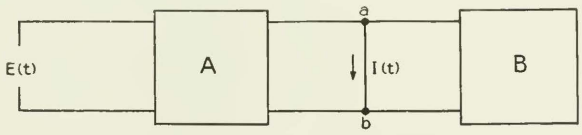


Fig. 32

Here the network is supposed to be energized by an electromotive force $E(t)$ which produced a current $I(t)$ in the *invariable* network in branch ab. We require the effect of suddenly opening this branch. The solution of this problem depends on the following proposition.

The effect of opening branch ab at reference time $t=0$ is the same as suddenly inserting at time $t=0$, a voltage $V(t)$ which produces in branch ab a current $-I(t)$ equal and opposite to the current which would exist in the branch in the absence of the open circuit.

While this proposition is precisely analogous to the corresponding proposition in the case of a sudden short circuit, it does not *explicitly* determine the voltage $V(t)$, which must be calculated as follows:

Let the driving point indicial admittance of the network, as seen from branch ab be denoted by $A_{ab}(t)$. Then, from the preceding proposition, it follows at once that $V(t)$ is given by

$$\frac{d}{dt}\int_0^t V(\tau)A_{ab}(t-\tau)d\tau = -I(t),\ t \geq 0.$$

This is a Poisson integral equation in $V(t)$, from which $V(t)$ is calculable. With $V(t)$ determined, the currents in any part of the network are calculable by usual methods, and consist of two components:—

(1) The current distribution in the network due to the impressed source $E(t)$ *in the absence of the open circuit.*

(2) The current distribution due to the electromotive force $V(t)$ inserted in branch ab at time $t=0$.

As in the case of the sudden short circuit, we have thus reduced the problem of a sudden open circuit to the determination of the current distribution in an *invariable* network in response to a suddenly impressed electromotive force.

Variable Circuit Elements

In the preceding cases of sudden open and short circuits it will be observed that the network changes discontinuously from one invariable state to another. A more general case, and one which includes the preceding as limiting cases, is presented by a network which includes a variable circuit element: that is, a circuit element which varies, continuously or discontinuously, with time. A network which includes such a variable circuit element will be called a *variable network.* Variable circuit elements of practical importance are the microphone transmitter, which consists of a variable resistance, varied by some source of energy outside the system; the condenser transmitter, which consists of a condenser of variable capacity; and the induction generator, in which the mutual inductance between primary and secondary, or stator and rotor, is varied by the motion of the latter. The case of a variable resistance will serve as an introduction to the general theory of such variable networks.

Consider a network, energized by a source $E(t)$ in branch 1, and containing a variable resistance element $r(t)$ in branch n. The functional notation $r(t)$ indicates that the resistance r varies with time. Let $I_n(t)$ denote the current in branch n, and assume that the network

is in equilibrium prior to the reference time $t=0$. The mathematical theory of this network depends on the following proposition:—

The network described above can be treated as an invariable network by eliminating the variable resistance element $r(t)$ and inserting an electromotive force $-r(t)I_n(t)$: that is, an electromotive force equal and opposite to the potential drop across the variable resistance element. Consequently the current in the variable resistance branch is determined analytically by the integral equation

$$I_n(t) = \frac{d}{dt}\int_0^t E(\tau)A_{1n}(t-\tau)d\tau - \frac{d}{dt}\int_0^t r(\tau)I_n(\tau)A_{nn}(t-\tau)d\tau. \qquad (288)$$

The first component is simply the current $I_o(t)$ which would exist in the variable branch if the variable element were absent; hence, dropping the subscript n for convenience, the current in the variable branch is given by the integral equation

$$I(t) = I_o(t) - \frac{d}{dt}\int_0^t r(\tau)I(\tau)A(t-\tau)d\tau \qquad (289)$$

and the voltage across the variable element by

$$v(t) = r(t)I(t). \qquad (290)$$

Having determined $I(t)$ and $v(t)$ from this integral equation, the distribution of currents in the network is calculable as that due to a source $E(t)$ in branch 1 and a source $v(t)$ in branch n of the *invariable network*: that is, the network with the variable resistance element eliminated.

A very simple example will serve to illustrate the foregoing:—

Into a circuit of unit resistance, and inductance $L=1/a$, in which a steady current I_o is flowing, a resistance r is suddenly inserted at time $t=0$: required the resultant current $I(t)$. In this case we have:

$$A(t) = \text{indicial admittance of unvaried circuit}$$
$$= 1-e^{-at},$$
$$r(t) = r,$$

and the integral equation of the problem is:

$$I(t) = I_o - r\frac{d}{dt}\int_0^t (1-e^{-ay})I(t-y)dy$$
$$= I_o - ra\int_0^t I(t-y)e^{-ay}dy.$$

If the solution is carried out as indicated by (291) below, and if the notation $at=x$ is introduced, we get without difficulty

$$I(t)=I_o\{1-r(1-e_1(x)e^{-x})+r^2(1-e_2(x)e^{-x})-r^3(1-e_3(x)e^{-x})+\ \ldots\},$$

where the function $e_n(x)$ is defined as:

$$e_n(x)=1+x/1!+x^2/2!+x^3/3!+\ \ldots\ +x^{n-1}/(n-1)!$$

$$=\text{first } n \text{ terms of the exponential series.}$$

For all finite values of the resistance increment r the series can be summed by aid of the identity

$$1-e_n(x)e^{-x}=\int_0^x dxe^{-x}x^{n-1}/(n-1)!$$

Substitution of this identity gives

$$I(t)=I_o\left(1-r\int_0^x e^{-(1+r)x}dx\right)$$

$$=I_o\frac{1+re^{-(1+r)x}}{1+r}.$$

Equation (289) is an integral equation of the Volterra type, which includes the Poisson integral equation as a special case. Its formal series solution is obtained as follows:—Assume a series solution of the form

$$I_n(t)=I_o(t)-I_1(t)+I_2(t)-I_3(t)+\ \ldots \tag{291}$$

and define the terms of the series by the scheme

$$I_1(t)=\frac{d}{dt}\int_0^t r(\tau)I_o(\tau)A(t-\tau)d\tau$$

$$\text{- - - - - - - - - - - - - - - - - - -} \tag{292}$$

$$I_{k+1}(t)=\frac{d}{dt}\int_0^t r(\tau)I_k(\tau)A(t-\tau)d\tau.$$

Direct substitution shows that this series satisfies the integral equation. Furthermore, it is easily shown that it is absolutely convergent.

While this series solution is not, in general, well adapted for numerical calculations, it throws a good deal of valuable light on the ultimate character of the oscillations in the important case where $E(t)$ and $r(t)$ both vary sinusoidally with time. In this case, if the frequency of the applied e.m.f. be denoted by F and that of the resistance variation by f, it is easy to show that the current $I_o(t)$ in the unvaried

circuit is ultimately [10] a steady state current of frequency F. This follows from the fact that the definite integral which defines the current $I_o(t)$ is resolvable into the ultimate steady state current corresponding to an applied force of frequency F, and the accompanying transient oscillations which ultimately die away. The fictitious e.m.f. which may be regarded as producing the component current $I_1(t)$ is $rf(t)I_o(t)$; this is ultimately the product of the two frequencies F and f, and therefore resolvable into two terms of frequency $F+f$ and $F-f$ respectively. Carrying through this analysis, it is easy to show that each component current is ultimately a steady-state but poly-periodic oscillation, as indicated in the following table:

Component Current	Frequency
I_0	F,
I_1	$F+f, F-f$,
I_2	$F+2f, F, F-2f$,
I_3	$F+3f, F+f, F-f, F-3f$,
I_4	$F+4f, F+2f, F, F-2f, F-4f$.

It is of importance to observe that the component currents involve, from a mathematical standpoint, multiple integrals of successively higher orders, the nth component $I_n(t)$ involving a multiple integral of the n^{th} order with respect to $I_o(t)$. Consequently the successive currents require longer and longer intervals of time to build up to their proximate steady-state values, so that the time required for the resultant steady-state to be reached cannot be inferred from the time constant of the unvaried circuit.

From the preceding table it will be seen that the ultimate steady-state current is obtained by rearranging the series $I_o+I_1+I_2$ and is of the form

$$\sum_{n=-\infty}^{+\infty} A_n \cos(\Omega+n\omega)t+B_n \sin(\Omega+n\omega)t$$

where $\Omega=2\pi F$ and $\omega=2\pi f$.

It is interesting to note that this series comes within the definition of a Fourier series only when $F=0$ or an exact multiple of f. The steady-state solution is of very considerable importance and is considered in more detail in a succeeding chapter.

From the foregoing we deduce an outstanding distinction between the variable and invariable networks. In the latter the currents are

[10] It hardly seems necessary to remark that the reference time $t=o$ is purely arbitrary and that the resistance variation may start at such a time thereafter that $I_o(t)$ may be regarded as steady state during the entire time interval in which we are interested. Going farther, if we confine our attention to sufficiently large values of t, the whole process may be treated as steady state.

ultimately of the same frequency as the impressed e.m.f., whereas in the former they are ultimately of an infinite series of frequencies.

In the preceding example, the variable impedance element is a resistance $r(t)$. If the variable element is taken as an *inductance* $\lambda(t)$ the voltage, corresponding to equation (290) is

$$\frac{d}{dt}\lambda(t)I(t).$$

The case of a variable capacity element is handled as follows: Let $1/C=S$ and assume that S is variable: thus, $S=S_0+\sigma(t)$. The drop across the variable condenser element is then

$$v(t)=\sigma(t)\int_0^t I(t)dt.$$

Similarly a variable mutual inductance $\mu(t)$ between branches m and n produces the voltages

$$\frac{d}{dt}\mu(t)I_n(t)$$

in branch m, and

$$\frac{d}{dt}\mu(t)I_m(t)$$

in branch n. This case may be illustrated by:

The Induction Generator Problem

In a sufficiently general form, this problem, which includes the fundamental theory of the dynamo, may be stated as follows:

Given an invariable primary and secondary circuit with a variable mutual inductance $Mf(t)$ which is an arbitrary but specified time function, and let the primary be energized by an e.m.f. $E(t)$ impressed in the circuit at the reference time $t=0$: required the primary and secondary currents.

In operational notation the problem may be formulated by the equations:

$$Z_{11}I_1-pMf(t)I_2=E(t),$$
$$-pMf(t)I_1+Z_{22}I_2=0,$$

in which Z_{11} and Z_{22} are the self impedances of the primary and secondary respectively; $Mf(t)$ is the variable mutual inductance; $E(t)$ is the applied e.m.f. in the primary; and p denotes the differential

operator d/dt. By aid of the fundamental formula these equations may be written down as the following simultaneous integral equations:

$$I_1(t) = \frac{d}{dt}\int_0^t dy A_{11}(t-y)\left(E(y) + M\frac{d}{dy}[f(y)I_2(y)]\right)$$

$$I_2(t) = M\frac{d}{dt}\int_0^t dy A_{22}(t-y)\frac{d}{dy}[f(y)I_1(y)].$$

In these equations, $A_{11}(t)$ and $A_{22}(t)$ denote the indicial admittances of the primary and secondary circuits respectively (when $M=0$): that is, the currents in these circuits in response to a unit e.m.f. (zero before, unity after time $t=0$). We assume, of course, that they are known or can be determined by usual methods.

It follows at once that the formal solution of these equations is the infinite series:

$$I_1(t) = X_0(t) + X_2(t) + X_4(t) + \ldots + X_{2n}(t) + \ldots$$

$$I_2(t) = Y_1(t) + Y_3(t) + Y_5(t) + \ldots$$

in which the successive terms of the series are defined as follows:

$$X_0(t) = \frac{d}{dt}\int_0^t dy A_{11}(t-y)E(y) = I_0(t),$$

$$Y_1(t) = M\frac{d}{dt}\int_0^t dy A_{22}(t-y)\frac{d}{dy}[f(y)X_0(y)],$$

$$X_2(t) = M\frac{d}{dt}\int_0^t dy A_{11}(t-y)\frac{d}{dy}[f(y)Y_1(y)],$$

$$Y_3(t) = M\frac{d}{dt}\int_0^t dy A_{22}(t-y)\frac{d}{dy}[f(y)X_2(y)], \quad \text{etc.}$$

In the light of formula

$$I(t) = \frac{d}{dt}\int_0^t f(y)A(t-y)dy$$

the physical interpretation of the series solutions follows at once: Thus, $X_0(t)$ is equal to the current $I_0(t)$ flowing in the *isolated* primary in response to the applied e.m.f. $E(t)$; the first component current $Y_1(t)$ in the secondary is equal to the current which would flow in the isolated secondary in response to the applied e.m.f. $M(d/dt)f(t)X_0(t)$; $X_2(t)$, the second component current in the primary, is equal to the current in the isolated primary in response to the applied e.m.f. $M(d/dt)$

$f(t)\,Y_1(t)$; etc. The resultant currents are thus represented as built up by a to-and-fro interchange of energy between primary and secondary, or by a series of successive reactions. In the important case where the applied e.m.f. and the variation of mutual inductance are both sinusoidal time functions, of frequency F and f respectively, it is easy to show that each component current becomes ultimately equal to a set of periodic steady-state currents. Thus the component X_o is ultimately single periodic, of frequency F; Y_1 is ultimately doubly periodic, of frequencies $F+f$ and $F-f$; X_2 triply periodic, of frequencies $F+2f$, F and $F-2f$; Y_3 quadruply periodic, of frequencies $F+3f$, $F+f$,$F-f$,$F-3f$; etc.

The Solution for the Steady-State Oscillations

For the very important case of periodic applied forces and periodic variations of circuit elements we are often concerned exclusively with the ultimate steady-state of the system, and not at all with the mode in which the steady-state is approached: that is, attention is restricted to the periodic oscillations which the system executes after transient disturbances have died away. In this case, if the periodic variations of circuit elements are sufficiently small, the required steady-state is obtained in the form of a series by replacing each term of the complete series solution by its ultimate steady-state value; a process which is very simple in view of the physical significance of each term of the latter series. The appropriate procedure will be briefly illustrated in connection with the variable resistance element. In view of the fact that we are concerned only with the ultimate steady-state oscillations, we can base the solutions on the symbolic equation

$$I = I_o - \frac{r(t)}{Z} I. \tag{293}$$

Here $r(t)$ is the variable resistance element; I_o is the current which would flow in the absence of the resistance variation; and Z is a generalized impedance of the network, as seen from the variable branch. Its precise significance and functional form is given below.

We now suppose that I_o is given by

$$I_o = J_o e^{i\Omega t} \qquad \text{(real part)} \tag{294}$$

$$= \tfrac{1}{2}\,(J_o e^{i\Omega t} + \bar{J}_o e^{-i\Omega t}) \tag{295}$$

where the bar indicates the conjugate imaginary of the unbarred

symbol, so that (295) is entirely real. Correspondingly the variable resistance will be taken as

$$\begin{aligned} r(t) &= \frac{r}{2}(e^{i\omega t}+e^{-i\omega t}) \\ &= re^{i\omega t} \qquad \text{(real part)} \\ &= r\cos\omega t. \end{aligned} \tag{296}$$

Here r is taken as a pure real quantity, which fixes the size of the resistance variation. No loss of generality is involved in this, since it merely involves referring the time scale to the zero of the resistance variation.

The symbolic impedance Z, as employed in the theory of alternating currents, will depend on the frequency and is, in general, a complex quantity. Its value at frequency $\Omega/2\pi$ will be denoted by

$$Z(i\Omega) = Z_o$$

while its value at frequency $(\Omega+n\omega)/2\pi$ will be written as

$$Z(i(\Omega+n\omega)\,) = Z_n.$$

We now assume a series solution of (293) of the form

$$I = I_o + I_1 + I_2 + \ldots$$

where the terms of the series are defined by the symbolic equations

$$\begin{aligned} I_1 &= -\frac{r(t)}{Z} I_o, \\ &\text{- - - - - - - - - - -} \\ I_{n+1} &= -\frac{r(t)}{Z} I_n. \end{aligned} \tag{297}$$

Substitution shows that this series formally satisfies the equation.

Starting with the first of (297) and substituting (295) and (296) we get

$$\begin{aligned} I_1 &= -\frac{r}{4Z}(e^{i\omega t}+e^{-i\omega t})\ (J_o e^{i\Omega t}+\bar{J}_o e^{-i\Omega t}) \\ &= -\frac{r}{4Z}\left\{ J_o e^{i(\Omega+\omega)t}+\bar{J}_o e^{-i(\Omega+\omega)t}+J_o e^{i(\Omega-\omega)t}+\bar{J}_o e^{-i(\Omega-\omega)t} \right\}, \end{aligned} \tag{298}$$

or

$$I_1 = -\frac{r}{2} J_o \left\{ \frac{e^{i(\Omega+\omega)t}}{Z_1} + \frac{e^{i(\Omega-\omega)t}}{Z_{-1}} \right\}. \tag{299}$$

In (299) it is to be understood that the real part is alone to be retained.

Proceeding in a similar way with the equation

$$I_2 = -\frac{r(t)}{Z} I_1$$

we get

$$I_2 = \left(\frac{r}{2}\right)^2 J_o \left\{ \frac{e^{i(\Omega+2\omega)t}}{Z_1 Z_2} + \frac{e^{i(\Omega-2\omega)t}}{Z_{-1} Z_{-2}} + \frac{e^{i\Omega t}}{Z_o}\left(\frac{1}{Z_1}+\frac{1}{Z_{-1}}\right) \right\}. \tag{300}$$

In this way the steady-state series solution is built up term by term, the component currents being poly-periodic as indicated in a previous table.

For sufficiently small impedance variations this method of solution works very well, and leads to a rapidly convergent solution. In other cases, however, the solution so obtained may be divergent, even when the complete series solution from which it is derived is absolutely convergent. The explanation of this lies in the fact that the steady-state series so obtained is the *sum of the limits* (as t approaches infinity) of the terms of the complete series solution, whereas the actual steady-state is the *limit of the sum.* These are not in general equal; in particular the former may be and often is divergent when the latter is convergent.

In view of the foregoing considerations it is of importance to develop another method of investigating the steady-state oscillations which avoids the difficulties in the formal series solution. The following method has suggested itself to the writer and works very well in cases where the previous form of solution fails. It should be stated at the outset, however, that the absolute convergence of the solution to be discussed, while reasonably certain in all physically possible systems, has not been established by a rigorous mathematical investigation, which appears to present very considerably difficulties.

We start with the problem just discussed and, in view of the results of the formal series solution there obtained, assume a solution of the form:

$$I = \frac{1}{2} \sum_{-N}^{N} A_m e^{i(\Omega+m\omega)t} + \bar{A}_m e^{-i(\Omega+m\omega)t} \tag{301}$$

$$= \sum_{-N}^{N} A_m e^{i(\Omega+m\omega)t} \quad \text{(real part)}. \tag{302}$$

Here the series is supposed to extend from $m = +N$ to $m = -N$. Ultimately, however, N will be put equal to infinity. As before, the

bar indicates the conjugate imaginary of the unbarred symbol and (301) is therefore entirely real.

If we now substitute (301) in the symbolic equation (293) we get, by (295) and (296),

$$\frac{1}{2}\sum\left\{A_m e^{i(\Omega+m\omega)t}+\bar{A}_m e^{-i(\Omega+m\omega)t}\right\}=\tfrac{1}{2}J_o e^{i\Omega t}+\tfrac{1}{2}J_o e^{-i\Omega t}$$

$$-\frac{r}{2Z}(e^{i\omega t}+e^{-i\omega t})\sum\left\{A_m e^{i(\Omega+m\omega)t}+\bar{A}_m e^{-i(\Omega+m\omega)t}\right\}.$$

Simplifying this equation and dropping the conjugate imaginaries gives:—

$$\sum A_m e^{i(\Omega+m\omega)t}=J_o e^{i\Omega t}-\frac{r}{Z}\sum A_m e^{i(\Omega+(m+1)\omega)t} \qquad (303)$$

$$-\frac{r}{Z}\sum A_m e^{i(\Omega+(m-1)\omega)t}.$$

Finally, if we write

$$Z(i(\Omega+m\omega))=Z_m$$

and

$$r/Z_m=h_m \qquad (304)$$

and equate terms of the same frequency on the two sides of the equation, we get

$$\begin{aligned} A_N &= -h_N A_{N-1} \\ A_m &= -h_m(A_{m-1}+A_{m+1}) \qquad 0<|m|<N \\ A_o &= J_o-h_o(A_{-1}+A_1). \end{aligned} \qquad (305)$$

It will be observed that, by (305), starting with A_N each coefficient is determined in terms of the coefficient of the next lower index. Thus:

$$A_N=-h_N A_{N-1}$$

$$A_{N-1}=-h_{N-1}(A_{N-2}+A_N)$$

$$=-\frac{h_{N-1}A_{N-2}}{1-h_{N-1}h_N}.$$

Similarly

$$A_{N-2}=-\frac{h_{N-2}A_{N-3}}{1-h_{N-2}h_{N-1}}\cdot\frac{1}{1-h_{N-1}h_N}.$$

Continuing this process it is easy to show that, for positive indices (m positive),

$$A_m = -h_m C_m A_{m-1} \tag{306}$$

where C_m designates the continued fraction

$$\cfrac{1}{1-h_m h_{m+1}\ \cfrac{1}{1-h_{m+1}h_{m+2}\ \cdot\ \cdot\ \cdot\ \cfrac{1}{1-h_{N-1}h_N}}}$$

The procedure for the coefficient A_{-m} is precisely similar. For convenience we write $A_{-m}=A'_m$, $Z_{-m}=Z'_m$, and $r/Z_{-m}=h'_m$. In this notation we get by precisely similar procedure

$$A'_m = -h'_m C'_m A'_{m-1} \tag{307}$$

where C'_m designates the continued fraction

$$\cfrac{1}{1-h'_m h'_{m+1}\ \cfrac{1}{1-h'_{m+1}h'_{m+2}\ \cdot\ \cdot\ \cdot\ \cfrac{1}{1-h'_{N-1}h'_N}}}$$

We now put the index N equal to infinity and the continued fractions C_m and C'_m become infinite instead of terminating fractions.

Collecting formulas we now have

$$A_m = -h_m C_m A_{m-1}$$

$$A'_m = -h'_m C'_m A'_{m-1}$$

and

$$A_o = J_o - (h_o A_1 + h'_o A'_1)$$

whence

$$A_o = \frac{J_o}{1-h_o h_1 C_1 - h_o' h_1' C_1'}.$$

The coefficients are thus all determined in terms of J_o.

The practical value of this method of solution will depend, of course, on the rate of convergence of the continued fractions. While no rigorous proof has been obtained, it is believed that they are absolutely convergent for all physically possible systems, but this question certainly requires fuller investigation. Nevertheless any doubt regard-

ing the convergence of the solution need not prevent the use of the method in a great many problems where physical considerations furnish a safe guide. For example this method of solution, when applied to the problem of the induction generator, discussed above, leads to the usual simplified engineering theory of the induction generator and motor, besides exhibiting effects which the usual treatment either ignores or fails to recognize.

Non-Linear Circuits

In the previous examples discussed, the variations of the variable circuit elements are assumed to be specified time functions, which is the same thing as postulating that these variations are controlled by ignored forces which do not explicitly appear in the statement and equations of the problem. We distinguish another type of variable circuit element, where the variation is not an explicit time function but rather a function of the current (and its derivatives) which is flowing through the circuit. For example, the inductance of an iron-core coil varies with the current strength as a consequence of magnetic saturation. The equation of a circuit which contains such a variable element (provided it is a single valued function) may be written down in operational notation

$$ZI+\phi(I)=E(t),$$

or

$$ZI=E(t)-\phi[I(t)]. \tag{311}$$

In this equation Z is, of course, to be taken as the impedance of the invariable part of the circuit, the indicial admittance of which is denoted by the usual symbol $A(t)$.

Equation (311) may be interpreted as the equation of the current $I(t)$ in a circuit of invariable impedance Z when subjected to an applied e.m.f. $E(t)-\phi[I(t)]$; consequently, by aid of our fundamental formula, $I(t)$ is given by

$$I(t)=\frac{d}{dt}\int_0^t A(t-y)E(y)dy-\frac{d}{dt}\int_0^t A(t-y)\phi[I(y)])dy.$$

The first integral is simply the current in the invariable circuit of impedance Z in response to the applied e.m.f. $E(t)$; denoting this by $I(t)$, we have

$$I(t)=I_0(t)-\frac{d}{dt}\int_0^t A(t-y)\phi[I(y)]dy.$$

This is a *functional integral equation*, the solution of which is gotten

by some process of successive approximations. For example, provided the sequence converges, $I(t)$ is the limit as n approaches infinity of the *sequence*

$$I_o(t),\ I_1(t),\ I_2(t),\ \ldots,\ I_n(t),$$

where the successive terms of the sequence are defined by the relations:

$$I_1(t) = I_o(t) - \frac{d}{dt}\int_0^t A(t-y)\phi[I_o(y)]dy,$$

$$\text{- -}$$

$$I_{n+1}(t) = I_o(t) - \frac{d}{dt}\int_0^t A(t-y)\phi[I_n(y)]dy.$$

We shall not pursue the discussion of non-linear circuits further, in view of their mathematical complexity and their relatively specialized technical interest. The reader who is interested may, however, consult the writer's paper on Variable Electrical Systems,[11] for a fuller treatment of the subject.

CHAPTER XI

The Application of the Fourier Integral to Electric Circuit Theory

The application of Fourier's series in electrotechnics is a commonplace; the use of the Fourier integral, however, has largely remained in the hands of professional mathematicians. An outstanding distinction between the series and the integral, from which the greater power of the latter may be inferred, is that the series represents only a periodic regularly recurrent function, whereas the integral is capable of representing a non-periodic function: in fact all types of functions, subject to certain mathematical restrictions which are usually satisfied in physical problems.

Before taking up the application of the Fourier Integral to Electric Circuit Theory, we shall very briefly review the elementary mathematics of the series and integral; for a fuller treatment the reader is referred to Byerly, Fourier's Series and Spherical Harmonics.[12]

Consider a function $\phi(t)$, which in the region $0 \leq t \leq T$ is finite, single-

[11] Phys. Rev. Feb., 1921.

[12] In this chapter the Fourier Integral is approached from the view-point of its physical application and no completeness or rigour is claimed for the treatment. The mathematical theory of the Fourier integral is, of course, completely developed in treatises on the subject. The object of this chapter is merely to outline some of its applications.

valued and has only a finite number of discontinuities or of maxima or minima. In this region it can then be expressed as the Fourier series

$$\phi(t) = \tfrac{1}{2} A_0 + \sum_1^\infty \left\{ A_n \cos\left(\frac{2\pi n}{T} t\right) + B_n \sin\left(\frac{2\pi n}{T} t\right) \right\} \tag{312}$$

where

$$\begin{aligned} A_n &= \frac{2}{T} \int_0^T \phi(t) \cdot \cos\left(\frac{2\pi n}{T} t\right) dt, \\ B_n &= \frac{2}{T} \int_0^T \phi(t) \cdot \sin\left(\frac{2\pi n}{T} t\right) dt. \end{aligned} \tag{313}$$

An equivalent series is

$$\phi(t) = \tfrac{1}{2} F_0 + \sum_1^\infty F_n \cos\left(\frac{2\pi n}{T} t - \theta_n\right) \tag{314}$$

where

$$\begin{aligned} F_n &= \sqrt{A_n^2 + B_n^2}, \\ \theta_n &= \tan^{-1}(B_n / A_n). \end{aligned} \tag{315}$$

This expansion is valid in the region $0 \leq t \leq T$, irrespective of the form of the function elsewhere. Let us, however, assume that the function repeats itself in the period T: that is

$$\phi(t \pm kT) = \phi(t), \qquad k = 1, 2, 3 \ldots N.$$

Then the expansion represents the function in the region $-NT \leq t \leq NT$. Finally if N is made infinite, the function is truly periodic and the Fourier series represents it for all positive and negative values of time.

It follows from the foregoing that, if the Fourier series represents the function for all positive and negative values of time, the function must be periodic for all positive and negative values of time; otherwise the expansion is valid only over a restricted range of time.

Now let us suppose that $\phi(t)$ is non-periodic. For convenience, in connection with subsequent applications we shall suppose that it is zero for all finite *negative* values of time, that it converges to zero as $t \to \infty$, and that

$$\int_0^\infty \phi(t) \; dt$$

exists. Such a function obviously cannot be represented by the usual Fourier series for all finite positive and negative values of time; it

can be represented, however, by the limiting form assumed by the series as the fundamental period T is made infinite. That is, we can assume that the function is periodic in an infinite fundamental period and this will not affect the expansion for finite positive and negative values of time. Proceeding in this way and putting the fundamental period T equal to infinity in the limit, the Fourier series (314) becomes an infinite integral and we get

$$\phi(t) = \frac{1}{\pi}\int_0^{\infty} F(\omega) \cdot \cos\left(\omega t - \theta(\omega)\right) d\omega \tag{316}$$

where

$$F(\omega) = \left\{ \left[\int_0^{\infty} \phi(t) \cos \omega t \, dt\right]^2 + \left[\int_0^{\infty} \phi(t) \sin \omega t \, dt\right]^2 \right\}^{\frac{1}{2}} \tag{317}$$

and

$$\tan \theta(\omega) = \int_0^{\infty} \phi(t) \sin \omega t \, dt \div \int_0^{\infty} \phi(t) \cos \omega t \, dt. \tag{318}$$

This is the *Fourier integral* identity of the function $\phi(t)$ and is valid for all finite positive and negative values of time.

In physical applications, particularly those to electric circuit theory, it is often convenient to employ exponential instead of trigonometric functions. The required transformation follows easily from the relation

$$e^{i\theta} = \cos\theta + i \sin\theta, \qquad i = \sqrt{-1}.$$

Thus if we write $2\pi/T = \omega_o$ the Fourier series (312) is easily reduced to the form

$$\phi(t) = \sum_{-\infty}^{+\infty} F(in\omega_o) e^{in\omega_o t} \tag{319}$$

where

$$F(in\omega_o) = \frac{1}{T}\int_0^{T} \phi(\tau)\, e^{-in\omega_o \tau}\, d\tau. \tag{320}$$

In precisely similar manner the Fourier integral (316) can be written as

$$\phi(t) = \int_{-\infty}^{\infty} F(i\omega) \cdot e^{i\omega t} d\omega \tag{321}$$

$$= \frac{1}{2\pi}\int_0^{\infty} \phi(\tau) d\tau \int_{-\infty}^{\infty} e^{i\omega(t-\tau)} d\omega. \tag{322}$$

Applications to Electric Circuit Theory

Let us assume that at time $t=-NT$, an electromotive force $E(t)$, periodic in fundamental period T, is impressed on a circuit of complex impedance $Z(i\omega)$, where ω denotes 2π times the frequency. Required the resultant current I.

For values of $t>-NT$ the electromotive force (see formula (319)) can be expressed as the Fourier series

$$E(t)=\sum_{-\infty}^{\infty} F(in\omega_o)e^{in\omega_o t}$$

where

$$F(in\omega_o)=\frac{1}{T}\int_0^T E(\tau)e^{-in\omega_o\tau}d\tau.$$

The resultant current for $t>-NT$ is therefore

$$I=\sum_{-\infty}^{\infty}\frac{F(in\omega_o)}{Z(in\omega_o)}e^{in\omega_o t}+\left|\begin{array}{l}\text{transient oscillations}\\ \text{initiated at time}\\ t=-NT.\end{array}\right.$$

If we are concerned with the current for values of $t\geq 0$, and if NT is made sufficiently large, the initial transients will have died away and the *complete current* for $t\geq 0$, will be given by

$$I=\sum_{-\infty}^{\infty}\frac{F(in\omega_o)}{Z(in\omega_o)}e^{in\omega_o t}. \tag{323}$$

This formula implies the periodic character of $E(t)$ for sufficiently large negative values of time. If, however, $E(t)$ is zero for negative values of time, we can employ the Fourier integrals (321) and (322) in precisely the same way and get, as the *complete* expression for the current for positive or negative values of time:—

$$I=\int_{-\infty}^{\infty}\frac{F(i\omega)}{Z(i\omega)}e^{i\omega t}d\omega \tag{324}$$

$$=\frac{1}{2\pi}\int_0^{\infty}E(\tau)d\tau\int_{-\infty}^{\infty}\frac{e^{i\omega(t-\tau)}}{Z(i\omega)}d\omega. \tag{325}$$

The infinite integrals (324) *and* (325) *formulate the current in the network, specified by the impedance function* $Z(i\omega)$, *in response to an electromotive force* $E(t)$ *impressed at time* $t=0$; *they therefore mathematically formulate, by aid of the Fourier integral identity, the fundamental problem dealt with in the preceding chapters and solved by aid of the operational calculus.*

No attempt will be made here to discuss the solution of the infinite integral (325), which is usually a problem presenting formidable difficulties, even to the professional mathematician. The general method of solution is by contour integration in the complex plane and the calculus of residues. By this process it has been successfully applied to the solution of special problems, and also to deriving some general forms of solution such as the expansion theorem solution.[13] Compared, however, with the operational calculus, it has no advantages from the standpoint of rigour, and lacks entirely the remarkable simplicity and directness of the Heaviside method.

In the direct solution of circuit problems, therefore, it is believed that the application of the Fourier integral is attended by few if any advantages, and presents formidable mathematical difficulties. On the other hand, there are certain types of problems encountered in circuit theory, where the Fourier integral is a powerful tool. These will be briefly discussed.

The Energy Absorbed from Transient Applied Forces

In many technical problems, the complete solution for the instantaneous current due to suddenly applied electromotive forces, although formally straight-forward, involves a prohibitive amount of labor. In yet others, the applied forces may be random and specified only by their mean square values. In such problems a great deal of useful information is furnished by the mean power and mean square current absorbed by the network, and to the calculation of these quantities, the Fourier integral is ideally adapted. Its application depends on the following proposition, due to Rayleigh (Phil. Mag., Vol. 27, 1889, p. 466), and its corollary.

Let a function $\phi(t)$, supposed to exist only in the epoch $0 \leq t \leq T$, be formulated as the Fourier integral

$$\phi(t) = \frac{1}{\pi} \int_0^{\infty} |f(\omega)| \cdot \cos [\omega t - \theta(\omega)] \, d\omega$$

where

$$f(\omega) = \left\{ \left[\int_0^T \phi(t) \cos \omega t \, dt \right]^2 + \left[\int_0^T \phi(t) \sin \omega t \, dt \right]^2 \right\}^{\frac{1}{2}}$$

$$\tan \theta(\omega) = \int_0^T \phi(t) \sin \omega t \, dt \div \int_0^T \phi(t) \cos \omega t \, dt.$$

[13] Bush, "Summary of Wagner's Proof of Heaviside's Formula." Proc. Inst. of Radio Engineers. Oct., 1917. Fry. "The Solution of Circuit Problems." (Phys. Rev. Aug., 1919).

Then

$$\int_0^T [\phi(t)]^2\, dt = \frac{1}{\pi}\int_0^\infty |f(\omega)|^2\, d\omega,$$

whereby the time integral is transformed into an integral with respect to frequency.

A corollary of this theorem is as follows:

If two functions $\phi_1(t)$, $\phi_2(t)$ supposed to exist only in the epoch $0 \le t \le T$, are formulated by the Fourier integrals

$$\phi_1(t) = \frac{1}{\pi}\int_0^\infty |f_1(\omega)| \cdot \cos\,[\omega t - \theta_1(\omega)]\, d\omega,$$

$$\phi_2(t) = \frac{1}{\pi}\int_0^\infty |f_2(\omega)| \cdot \cos\,[\omega t - \theta_2(\omega)]\, d\omega,$$

then

$$\int_0^T \phi_1(t)\phi_2(t)dt = \frac{1}{\pi}\int_0^\infty |f_1(\omega)| \,.\, |f_2(\omega)| \,.\, \cos(\theta_1 - \theta_2)d\omega.$$

The applications of these theorems to circuit theory proceeds as follows:—

If an electromotive force $E(t)$, supposed to exist only in the epoch $0 \le t \le T$, is applied to a network of complex impedance $Z(i\omega) = |Z(i\omega)|\, e^{i\beta(\omega)}$ we know from the preceding discussion of the Fourier integral, that the electromotive force $E(t)$ and current $I(t)$ are expressible as the Fourier integrals

$$\begin{aligned} E(t) &= \frac{1}{\pi}\int_0^\infty |f(\omega)| \,.\, \cos\,(\omega t - \theta(\omega))d\omega, \\ I(t) &= \frac{1}{\pi}\int_0^\infty \frac{|f(\omega)|}{|Z(i\omega)|}\cos(\omega t - \theta(\omega) - \beta(\omega))d\omega. \end{aligned} \tag{326}$$

It follows at once from Rayleigh's theorem that

$$\int_0^\infty I^2 dt = \frac{1}{\pi}\int_0^\infty \frac{|f(\omega)|^2}{|Z(i\omega)|^2}d\omega. \tag{327}$$

Now let I_n be the current absorbed in branch n; let $z(i\omega) = |z(i\omega)|\, e^{i\alpha(\omega)}$ be the impedance of that branch and let $E_n(t)$ be the potential drop across that branch. It follows at once from the corollary to Rayleigh's theorem that

$$W = \int_0^\infty E_n(t) I_n(t)dt = \frac{1}{\pi}\int_0^\infty \frac{|f(\omega)|^2}{|Z(i\omega)|^2}\,|z(i\omega)|\cos\,\alpha(\omega)d\omega. \tag{328}$$

Formulas (327) and (328) formulate the mean square current and mean power absorbed by the branch of the network under consideration, and enable us to calculate these quantities, even in the case of complicated networks, with a minimum of labor. Formula (327) is particularly well adapted to computation because the integrand is everywhere positive, permitting, in most problems, of easy numerical integration, whereas the analytical solution may be complicated.

Formulas (327) and (328) have been applied to the theory of selective circuits, to the problem of interference from random disturbances, including static, and to the theory of the Schrotteffekt. For the details of such applications, which will not be entered into here, the reader is referred to the following papers.

Transient Oscillations in Electric Wave Filters, Bell System Technical Journal, July, 1923.

Selective Circuits and Static Interference, Trans. A. I. E. E., 1924.

An Application of the Periodogram to Wireless (Burch & Bloehmsma), Phil. Mag., Feb., 1925.

The Theory of the Schrotteffekt (Fry), Journal Franklin Institute, Feb., 1925.

The Building-Up of Alternating Currents

Another application of the Fourier Integral, which may be briefly mentioned, is to the building-up of alternating currents in response to suddenly impressed sinusoidal electromotive forces. The investigation of this problem is of great importance to the communication engineer, since the excellence of a signal transmission system is to a considerable extent determined by the duration and character of the building-up phenomena.

In long transmission systems the calculation of the building-up current as a time function is extremely complicated and laborious if not practically impossible. Furthermore we are usually not concerned with the current as an instantaneous time function, but rather with its *envelope*. The envelope of the current can be formulated and calculated by modified Fourier integrals, by the following process.

Suppose that an e.m.f. $E \cos \omega t$ is suddenly applied, at reference time $t=0$, to a network of transfer impedance

$$Z(i\omega) = | Z(i\omega) | e^{iB(\omega)}.$$

The resultant current $I(t)$ may always be written as:

$$I(t) = \frac{1}{2}\frac{E}{|Z(i\omega)|}\left\{(1+\rho)\cos(\omega t - B) + \sigma \sin(\omega t - B)\right\}$$

$$= \frac{1}{2}\sqrt{(1+\rho)^2+\sigma^2}\,\frac{E}{|Z(i\omega)|}\cos(\omega t - B(\omega) - \theta)$$

where

$$\theta = \tan^{-1}\frac{\sigma}{1+\rho}$$

Evidently the functions ρ and σ, which it is our problem to determine, must be -1 and 0 respectively for negative values of t, and approach the limits $+1$ and 0, respectively, as $t \longrightarrow \infty$.

In an engineering study of the building-up process we are principally concerned with the *envelope* of the oscillations: hence with

$$\frac{1}{2}\sqrt{(1+\rho)^2+\sigma^2}.$$

Our problem is therefore to determine the functions ρ and σ and to examine the effect of the applied frequency $\omega/2\pi$ and of the characteristics of the circuit, on their rate of building-up and mode of approach to their ultimate steady values.

The functions ρ and σ can be formulated as the Fourier integrals

$$\rho = \frac{1}{\pi}\int_0^{\infty}[P_\omega(\lambda) + P_\omega(-\lambda)]\sin t\lambda \cdot \frac{d\lambda}{\lambda}$$

$$-\frac{1}{\pi}\int_0^{\infty}[Q_\omega(\lambda) - Q_\omega(-\lambda)]\cos t\lambda \frac{d\lambda}{\lambda}$$

$$\sigma = \frac{1}{\pi}\int_0^{\infty}[Q_\omega(\lambda) + Q_\omega(-\lambda)]\sin t\lambda \frac{d\lambda}{\lambda}$$

$$+\frac{1}{\pi}\int_0^{\infty}[P_\omega(\lambda) - P_\omega(-\lambda)]\cos t\lambda \frac{d\lambda}{\lambda},$$

where

$$P_\omega(\lambda) = \frac{|Z(i\omega)|}{|Z(i\omega)+i\lambda)|}\cdot\cos[B(\omega+\lambda) - B(\omega)],$$

$$Q_\omega(\lambda) = \frac{|Z(i\omega)|}{Z(i\omega+i\lambda)|}\cdot\sin[B(\omega+\lambda) - B(\omega)].$$

These formulas are directly deducible from the fact that the applied e.m.f., defined as zero for $t<0$ and $E\cos\omega t$ for $t\geq 0$, can itself be expressed as

$$\frac{E}{2}\cos\omega t\left[1+\frac{2}{\pi}\int_0^{\infty}\sin t\lambda\frac{d\lambda}{\lambda}\right].$$

For important types of transmission systems, including the periodically loaded line, these formulas have been successfully dealt with and solutions of a satisfactory approximate character obtained. For further details, the reader is referred to a paper on "The Building-Up of Sinusoidal Currents in Long Periodically Loaded Lines" (Bell System Technical Journal, October, 1924).

The foregoing must conclude our very brief account of the Fourier Integral and its applications in Electric Circuit theory; an adequate treatment of this subject would require a treatise in itself, and is beyond the scope of the present work. All that has been attempted is to give a very brief introduction to its significance in physical problems and a few of its outstanding applications in circuit theory. The reader who is interested in pursuing this subject further is referred to a paper by T. C. Fry on "The Solution of Circuit Problems" (Phys. Rev., Aug., 1919), which gives a rigorous discussion of the solution of the Fourier Integral by contour integration, together with some general forms of solution of the circuit problem.[1]

[1] It was planned to include in this paper a bibliography of the important papers bearing on the Heaviside operational method. This, however, has not been completed, but plans call for its publication in the next issue.—Editor.

Abstracts of Recent Technical Papers from Bell System Sources

Cipher Printing Telegraph Systems for Secret Wire and Radio Telegraphic Communications. G. S. VERNAM.[1] This paper describes a printing telegraph cipher system developed during the World War for the use of the Signal Corps, U. S. Army. This system is so designed that the messages are in secret form from the time they leave the sender until they are deciphered automatically at the office of the addressee. If copied while en route, the messages cannot be deciphered by an enemy, even though he has full knowledge of the methods and apparatus used. The operation of the equipment is described, as well as the method of using it for sending messages by wire, mail or radio.

The paper also discusses the practical impossibility of preventing the copying of messages, as by wire tapping, and the relative advantages of various codes and ciphers as regards speed, accuracy and the secrecy of their messages.

Methods of High Quality Recording and Reproducing of Music and Speech Based on Telephone Research. J. P. MAXFIELD and H. C. HARRISON.[2] The paper deals with an analysis of the general requirements of recording and reproducing sound, with the nature of the inherent limitations where mechanical records are used, and a detailed description of a solution involving, first, the use of electrical equipment for the purposes of recording and, second, the use of mechanical equipment based on electric transmission methods for reproducing.

Probably the most useful feature of the paper is the complete description of the application of electrical transmission theory to mechanical transmission systems. A detailed analysis is made of the analogies between the electrical and the mechanical systems.

Electrical and Photo-Electric Properties of Thin Films of Rubidium on Glass. HERBERT E. IVES and A. L. JOHNSRUD.[3] Films which spontaneously deposit on glass surfaces in a highly exhausted cell containing rubidium are electrically conducting, and photo-electrically active. A study of the photo-electric properties of a rubidium coated

[1] *A. I. E. E. Journal*, Vol. 45, pp. 109–115, Feb., 1926.
[2] *A. I. E. E. Journal*, Vol. 45, pp. 243–253, Mar., 1926.
[3] *Astrophysical Journal*, Vol. 52, pp. 309–319, Dec., 1925.

plane glass surface shows the normal and selective effects less well differentiated than for the similar coatings which form on metal plates. A rubidium film formed on the inside of a glass cylinder is found to exhibit, in the dark, a pure ohmic resistance. This decreases under illumination in a manner which appears to be explained as due to the liberation of photo-electrons which under a potential gradient form an added current along the tube.

The Influence of Temperature on the Photo-Electric Effect of 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 insure that any gases, if present, were condensed on the outer alkali metal coated walls. The temperature of the cathode was controlled by a stream of evaporating liquid air, whereby all temperatures between +20 and −180° C. could be attained and held constant and be measured. In these cells the variation of photoelectric current with temperature in sodium, potassium, and rubidium is continuous, without abrupt changes. 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 condition, smooth crystalline surfaces were obtained. With these annealed potassium surfaces, the variation of photoelectric 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 photoelectric effect is intimately connected with the strains produced in the surface by expansion and contraction with temperature.

Positive Rays in Thermionic Vacuum Tubes..[5] HERBERT E. IVES. Thermionic tubes in which a quantity of alkali metal is present exhibit not only the normal electron current from the heated filament, but a positive current, which at low filament temperatures may be

[4] *Journal of the Optical Society of America & Review of Scientific Instruments*, Vol. 11, No. 6, pp. 565–579, Dec., 1925.

[5] *Journal of the Franklin Institute*, Vol. 201, pp. 47–69, Jan., 1926.

many times larger than the negative current. The electron current is in general reduced by the positive rays, but at higher filament temperatures the reduction of space charge by the positive causes a considerable increase of the current over a limited voltage range. By immersing the tube in liquid air the positive ray effects are almost eliminated, indicating that the alkali metal vapor is the source of the rays, which are probably produced by contact of metal atoms with the hot filament.

A New Directional Receiving System.[6] H. T. Friis. Reduction of static interference, or to state it more correctly, reduction of the ratio of static to signal, has been, almost since the beginning of the radio art, the most important problem in radio engineering. It is now well known that static disturbances have definite points of origin and that the impulses which are detected at a receiving station have definite directions of propagation. A receiving system having no directional selectivity is, therefore, affected by static impulses from all directions and, in spite of many inventions, it has not yet been possible to improve its signal-static ratio except by limiting the frequency band transmitted. A system which, however, is so designed as freely to receive waves arriving from a limited range of directions is susceptible only to static disturbances propagated within that range, and large improvements in signal static ratio have been claimed for different types of directive antenna systems during the past few years.

A directional receiving system for radio telephony in which directional selectivity is obtained by combining the output voltages from two antennas is described in this paper. The main feature of the system is the arrangement for controlling the output voltages of the antennas, so that they may be combined to neutralize each other or to reinforce each other as desired. A double detection (super-heterodyne) receiver is employed and the output voltages, which are combined so as to produce the directional characteristic, are the intermediate frequency currents due to the waves received by the antennas and the beating oscillator currents. The control of these output voltages is effected by operating upon the beating oscillator currents.

High-Power Metallography—Some Recent[7] *Developments in Photomicrography and Metallurgical Research.* Francis F. Lucas. The

[6] Proceedings of the Institute of Radio Engineers, Vol. 13, No. 6, pp. 685–707, Dec., 1925.

[7] *Journal of the Franklin Institute*, Vol. 201, No. 2, pp. 177–216, Feb., 1926.

usual conception of high-power metallography seems to be great enlargement, indistinct definition and lack of resolution. Such results, generally, have been classed under the heading "empty magnification" because they have failed to show more detail than has been shown at lower magnifications and with objectives of less resolving ability. Oftentimes the pictures would be unintelligible taken by themselves, but the reason they are recognized at all is because the same structures have been seen and identified by low- or medium-power methods. Such high-power results are like an elastic band which has been stretched unduly. As the band is stretched it becomes more and more attenuated and finally snaps. If the optical image is stretched by enlargement the details of the image become less and less distinct and finally the image breaks down altogether, so that the detail and the background blend together into a hazy outline of what formerly was a sharp image.

High-power metallography as presented in this article consists of so preparing metallurgical specimens that crisp, brilliant images may be obtained and photographed at high powers and of achieving approximately the potential resolving possibilities of splendid objectives.

By improvements in the method of preparing metallurgical specimens and in the technique of manipulating the apparatus, "empty magnification" is no longer synonymous with high-power photomicrography.

It is the object of this contribution to show the application of this new tool for metallurgical research to the study of metal structures which heretofore have not been resolved and the nature of which has led to much speculation and to wide differences of opinion. A clear understanding of the current conceptions of magnification and resolution is essential and a knowledge of the limitations which were regarded for many years as restricting the employment of high powers will prove of value in the interpretation of the results obtained. For this reason a brief discussion follows which not only shows the method of approach in the present development, but indicates the path along which we may work to secure a higher order of resolution. By resolution is meant that property of a lens system which enables it to distinguish or "resolve" as separate and distinct units fine structural details spaced very close together.

Research and Engineering.[8] E. B. Craft. Research in industry—which the author mentions is of comparatively recent origin—is defined as the application of methods of systematic and logical deduc-

1 Address before the Engineers' Club, Phila., Oct., 1925. *Engs. and Engg.*, Jan. 1926, Vol. 43, pp. 11–19.

tions to our every-day industrial and technical problems. Such research necessarily is of a highly specialized nature and requires special training. What is equally important, as is pointed out by the author, is the need of properly organizing and directing this group of specialized workers. Since research is a creative process and hence particularly individualistic, one of the important problems in what the author calls "organized research" is the supplying of such an atmosphere that the worker realizes his own welfare and advancement to be adequately cared for in this system of group working. A number of examples of organized research are mentioned (radio and wire telephony, telephotography, ocean telegraphy, speech and hearing, artificial speech, phonograph recording and reproducing) as apropos to the point in question. The close relationship between engineering and research and the impossibility of the one getting along without the other is made clear. For the worker, there is pointed out the necessity of management and for those in charge the soundness of industrial research as a business proposition. Industrial research far from being a luxury has become a necessity.

Contributors to this Issue

THOMAS SHAW, S.B., Massachusetts Institute of Technology, 1905; American Telephone and Telegraph Company, Engineering Department, 1905–1919; Department of Development and Research, 1919—. Mr. Shaw's major activities have been devoted to development problems in loading telephone circuits, including the loading apparatus.

W. FONDILLER, B.S., College of the City of New York, 1903; E.E., Columbia University, 1909; M.A., 1913; Engineering Department, Western Electric Company, 1909–25; Bell Telephone Laboratories, 1925—. Mr. Fondiller is in charge of the General Development Laboratory and has been engaged in the design of loading coils and electric filters, and directing studies of the properties of materials and analysis of manual and machine switching apparatus design.

H. R. FRIIS, E.E., Royal Technical College in Copenhagen, 1916; Columbia University, 1919–20; Research Department, Western Electric Company, 1920–24; Bell Telephone Laboratories, 1925—. Mr. Friis' work has been largely in connection with radio reception methods and measurements. He has published papers on vacuum tubes as generators, radio transmission measurements and static interference.

RONALD M. FOSTER, S.B., Harvard, 1917; American Telephone and Telegraph Company, Engineering Department, 1917–19; Department of Development and Research, 1919—.

WALTER A. SHEWHART, A.B., University of Illinois, 1913; A.M., 1914; Ph.D., University of California, 1917; Engineering Department, Western Electric Company, 1918–24; Bell Telephone Laboratories, Inc., 1925—. Mr. Shewhart has been engaged in the study of the relationship between the microphonic and physicochemical properties of carbon.

HERBERT E. IVES, B.S., University of Pennsylvania, 1905; Ph.D., Johns Hopkins, 1908; assistant and assistant physicist, Bureau of Standards, 1908–09; physicist, Nela Research Laboratory, Cleveland, 1909–12; physicist, United Gas Improvement Company, Philadelphia, 1912–18; U. S. Army Air Service, 1918–19; research engineer, Western Electric Company, 1919–24; Bell Telephone Laboratories,

1925—. Dr. Ives' work has had to do principally with the production, measurement and utilization of light.

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 many papers on theory of electric circuits and electric wave propagation.

The Bell System Technical Journal

July, 1926

The Power of Fundamental Speech Sounds

By C. F. SACIA and C. J. BECK

Synopsis: This paper describes the continuing work on speech power by means of oscillographic studies of vowels, semi-vowels and consonants. A previous paper considered the characteristics of a few individual sounds from the power standpoint, but the principal emphasis was placed upon speech as a whole. In this later analysis, sounds are considered individually on the basis of instantaneous and mean power. A practical application of the results is suggested.

CONTINUING the work done on speech power by means of power oscillograms,[1] we have made additional reductions in the data relative to the vowels, semi-vowels and consonants and have also prepared a smaller amount of data on the power of the semi-vowels and the consonants from the amplitude oscillograms.[2] This is a preliminary study of the subject, at least in so far as the latter two classes of sounds are concerned, for these records of speech sounds were made to show all sounds in their true relative value hence the consonant sounds, being greatly inferior to the vowels were measurable to a correspondingly smaller degree of accuracy. We have gathered such data as the existing records could yield before future plans are completed to make a more comprehensive study of consonants.

Stop consonants are not so well characterized by the power data as are other types. The unvoiced stop consonants have two properties: a puff whose main frequency component is of the order of 50 cycles with a few ripples of high frequency; and a modifying effect upon the beginning or end of the vowel which immediately precedes or succeeds it. Hence, such a consonant is more of a controlling factor and lacks the essential properties of a discrete sound. In giving the data on the puff where it is measurable, we separate the low and high frequency components. In the case of the voiced stop consonants the vocal cord vibrations give the consonant more character of its own.

Mean Power and Peak Power

In the paper on speech power and energy, the "mean power," P_m, was derived (in the case of the vowel sounds) as the mean of the power taken throughout the interval of the vocal cycle. By the assumption of an appropriate arbitrary interval instead, say of the order of one

[1] B. S. T. J. Vol. IV No. 4. "Speech Power and Energy," by C. F. Sacia.
[2] B. S. T. J. Vol. IV No. 4. "Sounds of Speech," by I. B. Crandall.

one-hundredth of a second, the definition applies as well to consonant sounds and in addition has the same practical significance as that of the mean power of a vowel.

Mean power is thus a variable function of time, starting from zero, rising to a maximum and eventually falling to zero again as the sound is being uttered.[3] In studying an aggregate of speech sounds it is impracticable to have the final results in terms of these mean power curves; the most important discriminant of such a curve of any sound is its maximum ordinate, P_m. This value was used in the earlier study and has been given the name "syllabic power" when used in connection with the syllable as a whole. In the present case we shall abbreviate by simply calling it the "mean power of the sound." Similarly, when we are considering the consonant apart from the rest of the syllable we select the maximum value of P_m for that consonant.

Likewise, in considering the instantaneous power of a sound we select the height of the greatest peak occurring therein and for convenience we call it the "peak power."

All the averages hereinafter tabulated are the arithmetic averages of such maximum ordinates and not the integrated averages.

Normal and Conversational Values

We specify "normal" values as those derived from monosyllables spoken disconnectedly without accent but also without being slighted; while "conversational" values are derived from ordinary conversational speech. It does not follow that the arithmetic average of conversational values for a given sound should equal the average of the normal value, for the reason that some sounds are slighted much more frequently than others, as we shall see later.

The Consonants and Semi-Vowels

Of these sounds two independent sets of data are available: instantaneous peak power and mean power. The former is summarized in Table I. To explain the table in detail we take as an example the consonant, "t" as in "tap." There being one observation upon each of two speakers, the greatest observation showed 19 microwatts (peak) from the lips of the one speaker while the other speaker reached a peak of 13 microwatts, and the average of these two is 16. As in the paper on Speech Power and Energy, the corresponding values of power intensity in microwatts per square centimeter at the condenser transmitter are given in the group at the right. Since the relating factor is

[3] See "Speech Power and Energy," Fig. 1, page 628, for comparison of instantaneous and mean powers.

TABLE I

Normal Values of Peak Power in Microwatts for Two Speakers

(A) CONSONANTS

Consonant		Total from Voice			Per Cm² at Trans.		
Symbol	Key	Max.	Min.	Ave.	Max.	Min.	Ave.
b	bat	7	7	7	0.06	0.05	0.06
p	pot	7	6	6	0.06	0.05	0.05
*p	pot	128	0	64	1.04	0.	0.52
d	dot	7	1	4	0.06	0.01	0.04
t	tap	19	13	16	0.15	0.11	0.13
g	get	9	7	8	0.07	0.06	0.06
k	kit	9	4	6	0.07	0.03	0.05
dh	that	10	8	9	0.08	0.06	0.07
th	thin	1	0	1	0.01	0.	0.01
*th	thin	30	0	15	0.24	0	0.12
v	vat	29	21	25	0.23	0.17	0.20
*f	for	53	10	31	0.42	0.08	0.25
f	for	4	2	3	0.04	0.02	0.03
j	jot	26	23	24	0.21	0.19	0.20
ch	chat	61	43	52	0.49	0.35	0.42
zh	azure	53	23	38	0.43	0.19	0.31
sh	shot	133	97	115	1.08	0.79	0.93
z	zip	42	21	31	0.34	0.17	0.25
s	sit	54	8	31	0.43	0.06	0.25

* Low frequency puff.

(B) SEMI-VOWELS

Semi-Vowel		Total from Voice			Per Cm² at Trans.		
Symbol	Key	Max.	Min.	Ave.	Max.	Min.	Ave.
l	let	226	37	131	1.83	0.29	1.06
ng	ring	169	25	97	1.36	0.20	0.78
n	no	74	21	47	0.59	0.17	0.38
m	me	198	23	111	1.60	0.18	0.89

NOTE: For these two speakers, the peak power of the succeeding vowel was as follows:

	Total	Per Cm²
ū (tool)	206	1.7
á (tap)	860	6.8
ē (teem)	241	1.9

about 127, the intensities 0.15, 0.11 and 0.13 are the first three numbers respectively divided by 127.

These values were derived by measuring the amplitudes of the above-mentioned oscillograms of the acoustic pressure. The maximum or peak amplitudes of the consonant and the succeeding vowel were first measured; the square of the ratio between these is the ratio of the

corresponding peak powers. Now the approximate peak powers of these vowels for the two speakers were found (see note under Table I) from the power oscillograms used in our study of speech power. Hence from the product we derive the approximate peak power of the consonant (or semi-vowel). Direct measurement of peak power from the latter oscillograms was impracticable because of the low sensitivity of the instantaneous power recorder [4] and the before-mentioned fact that the power of the consonants and semi-vowels is low relative to that of the vowels.

Since frequencies of the order of 50 cycles are of negligible importance in speech, the 50-cycle puff has been separated from the other components in the case of the unvoiced stop consonants. This is justified by the fact that the utterances of such a sound by two speakers may seem exactly alike to the careful listener, whereas a large puff may be present in one case and none in the other.

The values thus far considered represent "normal" values in speech —not accented and yet not slighted.

TABLE II

Conversational Values of Mean Power in Microwatts for 16 Speakers

(A) Consonants

Consonant		Speaker's Power		Number of Measurable Observations	Per Cm² at Trans.	
Symbol	Key	Max.	Av.		Max.	Av.
d	dot	2.9	0.08	4	0.023	0.0006
t	tap	6.0	0.14	14	0.049	0.0012
k	kit	4.8	0.34	20	0.039	0.0027
v	vat	2.4	0.03	1	0.019	0.0002
f	for	3.6	0.08	1	0.029	0.0006
j	jot	3.6	0.47	8	0.029	0.0038
ch	chat	7.9	1.44	19	0.064	0.0116
sh	shot	6.0	1.83	9	0.049	0.0148
z	zip	7.2	0.72	31	0.058	0.0058
s	sit	8.7	0.94	115	0.070	0.0076

(B) Semi-Vowels

Semi-Vowel		Speaker's Power		Number of Measurable Observations	Per Cm² at Trans.	
Symbol	Key	Max.	Av.		Max.	Av.
l	let	9.6	0.33	13	0.078	0.0026
ng	ring	3.6	0.35	2	0.029	0.0028
n	no	18.0	2.11	146	0.145	0.0170
m	me	16.8	1.85	31	0.136	0.0149

[4] In recording the power, separate vibrators had been used for instantaneous and mean powers.

Our measurements of mean power, on the other hand, were made from power records of conversational speech, with a greater variety of observations and speakers. Stress, therefore, plays an important part here.

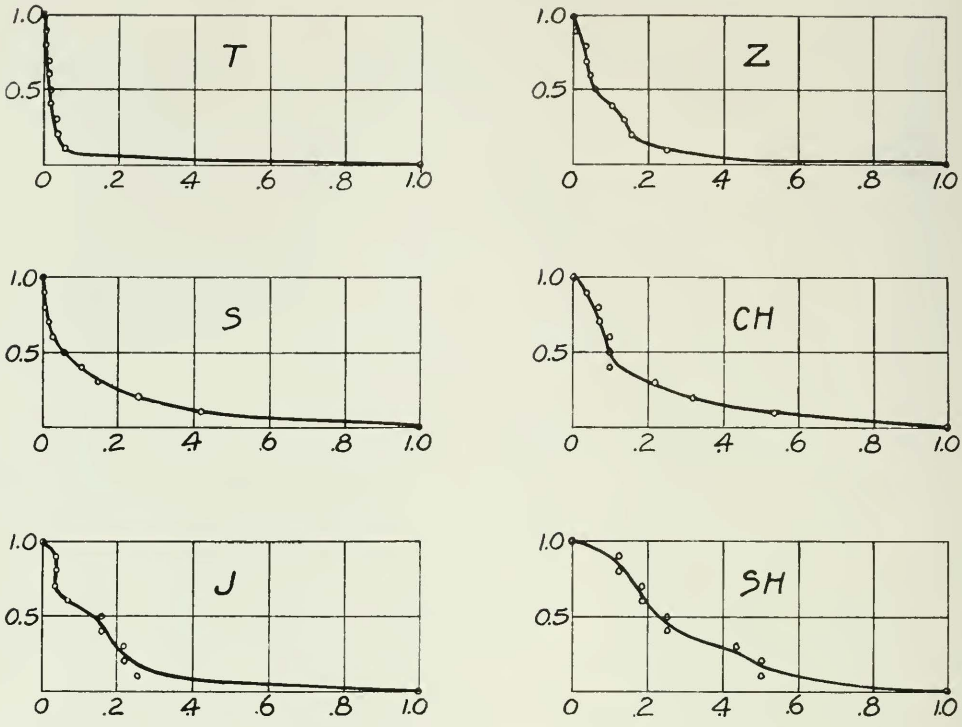
In Table II is given a compact summary of the direct measurements made on the power oscillograms. Thus consider "d" as in "dot." 2.9 microwatts was the greatest observed value for any speaker, while the average of all observations (including accented and unaccented utterances) was but 0.08. Only four observations, however, were large enough to be measured. As before, we give the corresponding intensities in microwatts per square centimeter at the transmitter in the next two columns.

To show the occurrence of stress in the utterance of these sounds in ordinary speech, we give in Fig. 1 the stress frequency-distribution curves [5] of several oft-occurring sounds. These curves are derived in the same manner as were the syllabic stress curves in the study of speech power. They exhibit the marked degree in which the consonants differ in stress for ordinary speech. For example, among the consonant sounds, "t" and "sh" represent extreme types. The former is either slighted or strongly accented with but little intermediate gradation while the blunt characteristic of the latter indicates the most nearly uniform distribution of stress into all shades from zero to maximum. Similarly with the three semi-vowels shown, "l" and "m" are extreme types.

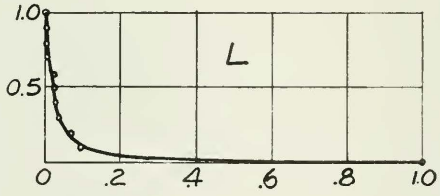
The Vowels

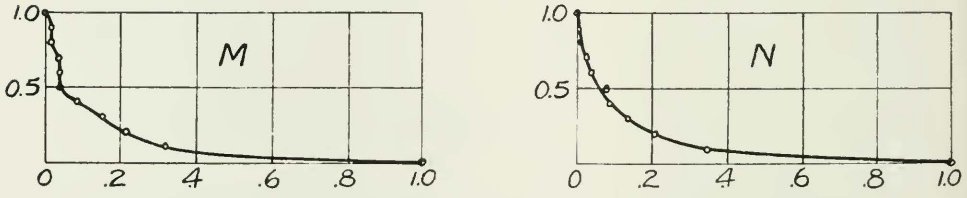
Some attention was given to vowel power in the other paper where under the heading of "Relative Power of Vowels" (on page 634) were charted what we have classified as normal values of mean power. These were derived from the mean power curves of disconnected monosyllables. Although they were charted separately for male and female voices, we shall not differentiate between the two in the following. In Tables III and IV are summarized the four sets of data based upon the speech from 16 voices. Here we see the influence of stress by comparing the conversational and normal values. This effect is noteworthy in the case of "o" (ton) "à" (tap) and "i" (tip) which average considerably less power in conversational speech than in normal syllables. Another point of interest is the comparison of peak and mean values. For example, in the normal data, the ratio of peak to mean (i.e. the

[5] The abscissa represents the relative number of observations (s/Σ) whose relative power values exceed the magnitude of the ordinate, n, a numeric varying between zero and one.



(A). CONSONANTS





(B). SEMIVOWELS

Fig. 1. Power Stress Curves.

TABLE III—VOWELS

Peak Power in Microwatts for 16 Speakers

Vowel		Total from Voice				Per Cm² at Trans.				Number of Measurable Observations Conversational Values
		Normal Values		Conversational Values		Normal Values		Conversational Values		
Symbol	Key	Max.	Av.	Max.	Av.	Max.	Av.	Max.	Av.	
ū	tool	620	290	760	180	5.0	2.3	6.1	1.5	61
u	took	890	470	——	——	7.2	3.8	——	——	0
ō	tone	1310	540	900	330	10.6	4.4	7.2	2.7	62
ò	talk	1240	630	1580	600	10.0	5.1	12.8	4.8	32
o	ton	1240	600	1720	300	10.0	4.8	13.9	2.4	248
a	top	1650	760	1580	660	13.3	6.1	12.8	5.3	127
à	tap	1860	1020	1380	290	15.0	8.2	11.1	2.3	38
e	ten	1720	660	1510	340	13.9	5.4	12.2	2.8	125
ā	tape	1380	580	1720	470	11.1	4.7	13.9	3.8	32
i	tip	1240	520	1330	190	10.0	4.2	10.8	1.5	198
ē	teem	1510	430	960	190	12.2	3.5	7.8	1.5	56
r̄	err	——	——	550	200	——	——	4.4	1.6	33

Note: The dash indicates that observations were not available.

TABLE IV—VOWELS

Mean Power in Microwatts for 16 Speakers

Vowel		Total from Voice				Per Cm² at Trans.				Number of Measurable Observations Conversational Values
		Normal Values		Conversational Values		Normal Values		Conversational Values		
Symbol	Key	Max.	Av.	Max.	Av.	Max.	Av.	Max.	Av.	
ū	tool	60	33	53	13	0.49	0.27	0.43	0.11	64
u	took	108	40	——	—	0.87	0.32	——	——	0
ō	tone	82	38	68	22	0.66	0.31	0.55	0.18	64
ò	talk	91	43	125	47	0.74	0.35	1.01	0.38	32
o	ton	84	33	107	15	0.68	0.27	0.86	0.13	284
a	top	111	48	130	34	0.89	0.39	1.05	0.28	128
à	tap	96	40	40	9	0.78	0.33	0.32	0.07	48
e	ten	79	27	88	17	0.64	0.22	0.71	0.13	141
ā	tape	58	26	62	20	0.47	0.21	0.49	0.16	32
i	tip	53	30	55	9	0.43	0.24	0.44	0.07	250
ē	teem	65	27	78	12	0.52	0.22	0.63	0.10	64
r̄	err	——	——	30	10	——	——	0.24	0.08	40

Note: The dash indicates that observations were not available.

square of the peak factor) is greater for centrally located vowels and is greatest for "à" (tap) as was mentioned in the earlier paper. Referring to the normal values of peak power we find a surprising degree of regularity in the increase of these values from a minimum for "ū" (tool) to a maximum for "à" (tap) and the falling off again to minimum for "ē" (teem). The one slight irregularity is the vowel "o" (ton). (We have omitted "r" (err) from this comparison because it has no well defined place on the Victor triangle which forms the basis for this arrangement of the other vowels).

TABLE V—SPEECH SOUNDS

Speech Sound	Key	Relative Power, Arbitrary Units		
		A Mean Power Conversational values for 16 speakers	B Peak Power Normal values for 2 speakers	C Relative Power Attenuation to give 80% Articulation
ò	talk	1870	688	826
a	top	1380	1430	474
ō	tone	875	630	619
ā	tape	808	632	567
e	ten	664	975	364
o	ton	616	688	474
ū	tool	532	344	349
ē	teem	484	402	421
r̄	err	384	– see note	924
à	tap	366	2170	645
i	tip	346	688	295
n	no	84	78	36
m	me	74	185	38
sh	shot	73	192	216
ch	chat	58	87	64
s	sit	38	51	11
z	zip	29	52	17
j	jot	19	41	98
ng	ring	14	162	134
k	kit	14	10	43
l	let	13	218	157
t	tap	6	26	32
d	dot	3	7	60
f	for	3	6	9
v	vat	1	41	13
u	took	– see note	688	347
zh	azure	–	63	–
dh	that	–	15	–
g	get	–	13	60
b	bat	–	11	30
p	pot	–	11	24
th	thin	–	1	1

NOTE: The dash indicates that observations were not available.

Relative Power of Speech Sounds

A direct comparison of most of the fundamental sounds will now be made. In Table V—A are shown the conversational values (averaged) of the mean power for each sound for 16 speakers. The units are taken arbitrarily in order to show only the relative values. As might have been expected, the vowels rank the highest, the semi-vowels next and the consonants the lowest, although we find a few consonants interspersed among the semi-vowels. In Table V—B is the similar arrangement for the normal values of peak power for the two speakers. Data on a larger number of sounds are available for this group, but the same general order prevails: vowels, semi-vowels and consonants. Minor differences in order (note "v" as in "vat") may be expected to occur because of the influence of stress upon the conversational value. But in both cases the ratio of the maximum to the minimum is of the order of 2000. This similarity is striking in view of the difference in the modes of utterance and the numbers of speakers in the two cases.

Finally, in Table V—C are shown relative values[6] derived on the basis of relative attenuation in power required to bring the articulation (as judged by the average ear) to 80%. Since disconnected monosyllables

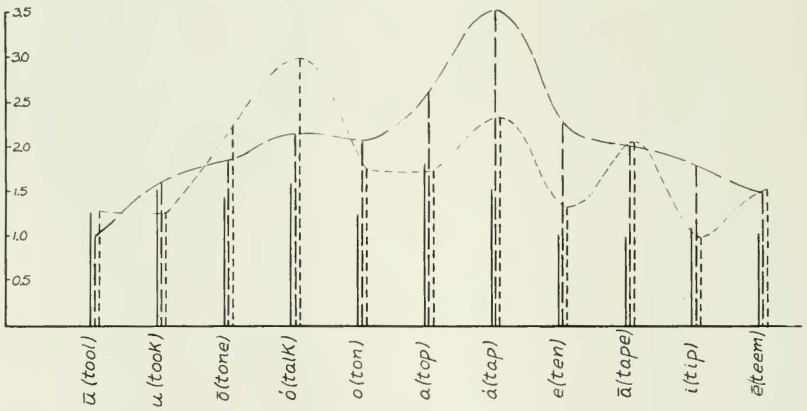


Fig 2. Comparative Chart Relative Normal Values of Vowel Sounds.
— — — — —Peak Power.
————————Mean Power.
– – – – – – – – –Relative Power Attenuation Required to Give 80% Articulation.

were used in this test the values are normal values in our present category. Although the same general order of the other two tables

[6] Taken from the paper presented by Harvey Fletcher before the Modern Languages Association, December 1923. Values are there called relative "intensity" which term we avoid here because of the acoustic meaning already assigned to intensity: power per square centimeter.

prevails here, there are considerable differences throughout which may well be expected since the ear is used in making the balance. The frequency response characteristic of the ear is the complicating factor in this case. The ratio of maximum to minimum here is of the order of one thousand or about one-half the absolute power ratio found in the two preceding tables.

A more orderly comparison between power and "relative attenuation" exists in the case of the vowels alone as shown in the chart of Fig. 2. Thus the peak power and "relative attenuation" most nearly correspond at the ends of the chart (especially the left) where there is resonance of lower frequency in the vowels. The vowel "o" again shows a peculiarity in that the two trends—as shown by the envelopes —intersect here. Peak power predominates over "relative attenuation" in the three successive vowels "a," "à," "e," which have strong resonance in the region from 600 to 1200 cycles. The vowel "i" gives the only erratic turn in this comparison, differing considerably from the two adjacent vowels.

As for loudness in the ordinary sense, let us note a phenomenon of rather common occurrence in these days of good quality sound reproducing apparatus. One may be listening to well reproduced speech at ordinary volume when suddenly a slightly accented syllable containing "à" (tap) comes through with noticeable overload distortion and its accompanying disagreeable effect upon the ear. Although the listener does not judge this sound to be any louder than numerous accented sounds preceding and following it, still the fact remains that there has been considerable overload due to the peaks of the wave being cut off by the amplifier. Where do we look for the explanation? As noted in the earlier paper this vowel has the highest peak factor, and we have already seen in Table III that it normally contains the greatest peak power. In spite of this therefore, it would seem that the loudness of this sound does not predominate over the loudness of the sounds in the first half of the chart, as does the peak power. This phenomenon can also be demonstrated, for the vowel "ē" (teem) and to a lesser degree even for the vowels which intervene between these two in the tables and chart of the vowel sounds.

Extraneous Interference on Submarine Telegraph Cables

By J. J. GILBERT

Synopsis: In order to avoid a considerable reduction in speed of operation, which would have resulted on account of the unusually large parasitic disturbances encountered in the neighborhood of New York, the New York-Azores permalloy loaded cable was equipped with a new type of earth connection consisting of a conductor extending 100 nautical miles to sea and there connected to earth through an artificial line.

This paper presents the theory of the new type of sea earthing arrangement and discusses the sources of extraneous interference and the manner in which it is picked up by submarine cables. A method is developed for estimating the magnitude of terminal extraneous interference in the case of any particular cable.

AMONG the factors limiting the speed of operation of long submarine telegraph cables one of the most important is the mutilation of the received signals by electrical disturbances picked up along the cable and transmitted with the incoming signal to the receiving instrument. The nature of this disturbance is shown in Fig. 1 which is an oscillographic record over a short period of time of the difference of potential across the terminals of the receiving instrument of a cable system, at a time when no signals were being received over the cable. Although the complete signal correction networks were not in circuit at the time this record was taken, the latter is representative of the form of the extraneous disturbance that would be superposed on the record of an incoming signal. It is evident that unless the signal amplitude is sufficiently large compared with the amplitude of interference, the latter will seriously interfere with the interpretation of the siphon recorder tape or with the functioning of relays operated by the signal current. That this condition constitutes a limit on the speed of operation of the cable is indicated by Fig. 2 which shows the amplitude of a signal, received over a typical transatlantic cable, as a function of

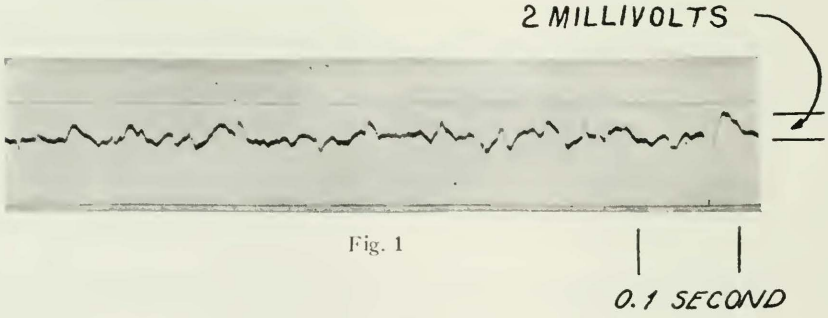
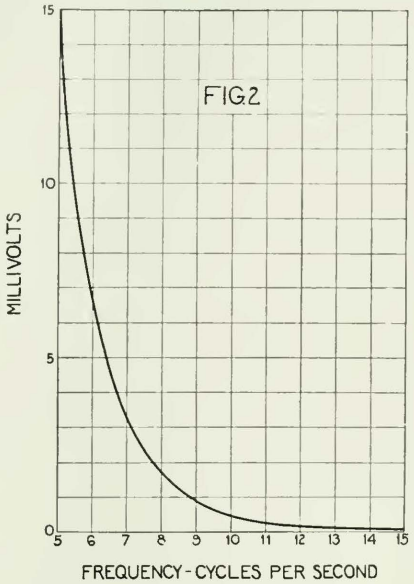


Fig. 1

the signal frequency.[1] It is evident, that corresponding to the minimum amplitude at which signals are just legible through interference, there is, for a given value of sending voltage, a maximum speed of signalling which cannot be exceeded, without danger of serious mutilation of the signal. If by any means the magnitude of the extraneous interference can be diminished, signals of smaller amplitude can be employed and the speed of operation consequently increased.



The present paper will be devoted to a description of the manner in which extraneous interference is picked up by submarine cables, with a discussion of the influence of various factors such as depth of water, cable structure and operating conditions. There will also be described a method of reducing interference by a modification of the cable structure. This method has been remarkably successful in the case of the New York-Azores continuously loaded cable,[2] and has helped to make available the great gain in operating speed due to continuous loading, which is the outstanding feature of this cable installation.

The disturbances encountered on submarine cables are due mainly to induction from extraneous electromagnetic fields in the sea water,

[1] The signal frequency is defined as the fundamental frequency involved in a succession of alternately positive and negative unit impulses.

[2] Buckley, O. E., *Journal A. I. E. E.*, Vol. XLIV, p. 821, August 1925, *Bell System Technical Journal*, Vol. IV, No. 3, July 1925.

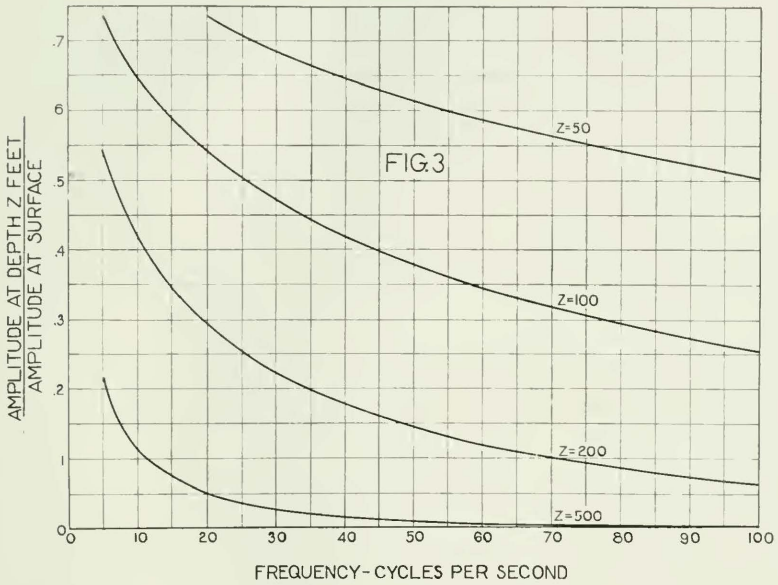
arising from a variety of sources, which may be broadly grouped into two classes. The first class comprises artificial sources, such as electrical power or railway systems in the neighborhood of the cable terminals. Currents circulating between the various earth connections of such systems give rise to electromagnetic fields in the earth and sea water, which fields may have the form of transient surges or pulses, or may be periodic in nature. The second class includes the various manifestations in the atmosphere or at the surface of the earth, such as electric or magnetic storms, which are also responsible for the disturbances in radio communication known as "static." Very little definite data is available regarding the magnitude and character of the natural disturbances affecting submarine cables, but it is found that, as in the case of static, the intensity of such effects is influenced by a number of factors, the season of the year and the geographical location being among the most important. At times of unusual activity, such as that accompanying the aurora polaris or local electrical storms, the voltages induced in the cable conductor are so large as to prohibit operation of the cable.

Except in the case where the source is in the immediate vicinity of the cable, the effect of any disturbance upon the cable can be considered as the result of a fluctuation of potential at the surface of a massive conducting medium, the ocean, which gives rise to electromagnetic waves which are propagated in all directions from the source and which penetrate the interior of the conducting medium according to the well-known laws governing "skin effect." Due to the presence of varying electric and magnetic fields in the sea water adjacent to the cable, an electromotive force is induced in each section of the cable conductor, and the resulting current is transmitted along the conductor to the cable terminal, combining with the currents due to electromotive forces induced in other sections to make up the total extraneous interference.

At the surface of the ocean the disturbance may take a variety of forms, for instance a succession of pulses or a train of damped oscillations. In any case the most convenient method of following the disturbance through the sea water into the cable conductor and along the conductor to the cable terminal is to consider the disturbance made up of a number of sinusoidal components of all frequencies from zero to infinity, the relative amplitudes and phases of the various components being determinable from the wave shape of the disturbance by the methods of Fourier analysis. The transmission characteristics of the interference transmission system at any particular frequency can then easily be studied, and finally the total effect of the original dis-

turbance can be obtained by summation of disturbances of all frequencies.

The extent to which electrical disturbances penetrate below the surface of the ocean can be determined from the theory of induction of currents in continuous media, where it is shown that the components of the electric (E) and magnetic fields (H) parallel to and at a distance



z below the surface of an infinite plane conductor are given by the formulas:[3]

$$E = E_o\, \epsilon^{-kz},\ H = H_o\, \epsilon^{-kz},\ k = 2\pi\sqrt{2\lambda i f}, \qquad (1)$$

where E_o and H_o are the values of E and H at the surface, λ is the electrical conductivity of the medium and f is the frequency. Employing the value of λ for sea water and expressing z in feet, gives

$$k = 1.35 \times 10^{-3}\sqrt{f}\ (1+i).$$

The curves of Fig. 3, computed from formula (1), indicate the manner in which sinusoidal disturbances of frequencies in the telegraph range are attenuated by various depths of sea water. It can be seen that the magnitude of a disturbance falls off rapidly as it penetrates the water; also that this attenuating effect is greater the higher the frequency. At a depth of one or two miles, at which the greater part

[3] Jeans "Electricity and Magnetism," 2nd Edition, p. 477.

of the typical transoceanic cable is submerged, only the extremely low frequency components of the surface disturbance are encountered to an appreciable degree. In the vicinity of the terminals, however, where the water is comparatively shallow, the cable is exposed to the higher frequency components of the disturbances, and it is usually in these sections that the greater part of the most troublesome disturbances is picked up. This is especially true in localities where the zone of shallow water extends a considerable distance from shore. Such a case is shown in Fig. 4, which represents a typical profile of the ocean bottom for the shallow water portion of a cable terminating at New York.

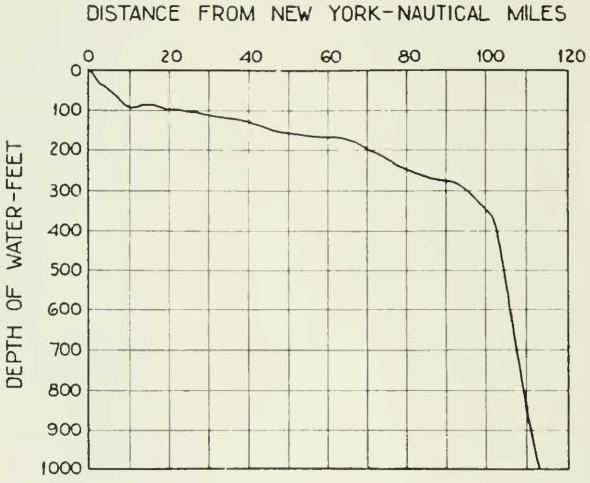


FIG. 4

The phenomena attending the induction of an electromotive force in the cable conductor by an electromagnetic field are rather complicated and difficult of exact computation. In the first place, on account of the change in electrical constants in passing from sea water to ocean bottom, the electric and magnetic field intensities in the neighborhood of the cable are somewhat different than indicated by equation (1). The influence of this factor upon the final result is in general small compared with that of the other factors that we are considering, and, on account of our lack of knowledge concerning the electrical characteristics of the ocean bottom, theoretical discussion would be of little practical value. A second factor is the shielding effect of the armor wires and metallic tapes surrounding the core. No attempt will be made in the present paper to work out an analytical solution of this

problem. There is available, however, from a recent study of the problem of the sea return resistance of a submarine cable,[4] information that enables us to compare the behavior of various cable structures from the point of view of shielding. One of the results of this work was the determination of the degree to which the shielding effect of the metallic sheath around the cable causes the returning signal current to flow in this sheath rather than in the surrounding sea water. It is obvious that the greater the tendency of the metallic sheath to confine the return current to itself, the more effective the sheath will be in reducing the pick-up of interference. Allowing for the two effects just discussed, it is evident that the electromotive force induced in unit length of the cable conductor is given by an expression of the form

$$e = A\ E_o\ \epsilon^{-kz} \tag{2}$$

where A is a multiplier, the value of which will be determined only on a relative scale.

The electromotive force induced in any section of the cable conductor gives rise to sinusoidal currents and potentials which are transmitted in both directions along the conductor in accordance with well-known laws. For simplicity we will assume that the cable is terminated at both ends in its characteristic impedance, Z, the result corresponding to any other values of terminal impedances being readily determinable if needed.[5] Then an electromotive force $e\,dx$, induced in a short section of cable of length dx, distant x from the terminal, will result in a current

$$\frac{e\,dx}{2Z}\,\epsilon^{-\gamma x} \tag{3}$$

at the terminal. If the electromotive force per unit length e is picked up uniformly over a length of cable extending from $x=a$ to $x=s$, then since the impedance in each direction from the point is Z, the resulting current at $x=o$ will be

$$\frac{e}{2Z}\int_a^{a+s} dx\ \epsilon^{-\gamma x} \tag{4}$$

$$= \frac{e}{2Z}\,\epsilon^{-a\gamma}\,\frac{1-\epsilon^{-s\gamma}}{\gamma}$$

[4] Carson and Gilbert "Transmission Characteristics of a Submarine Cable," *Jour. Franklin Inst.*, Vol. 192, p. 705, 1921, and *Electrician*, Vol. 88, p. 499, 1922; *Bell System Technical Journal*, Vol. I, No. 1, July 1922.

[5] Heaviside, "Electromagnetic Theory," Vol. 2, p. 75.

Thus the effect at $x=o$ is the same as if an electromotive force $e\frac{1-\epsilon^{-s\gamma}}{\gamma}$ had been impressed at $x=a$.[6]

It would now be possible to assume a definite form of disturbance at the surface of the ocean, and by applying the principles that have been discussed in the preceding pages, to work out for any particular cable the wave shape of the resulting interference at the cable terminals. On account of our lack of knowledge as to what might be considered a typical disturbance at the surface of the ocean, such results would be merely speculative, and would be of no practical value in predicting the actual terminal interference that might be expected. A much better scheme is to compute for each cable, what may be called the interference susceptibility, this being defined, for a particular frequency, as the integral

$$\int A \cdot \epsilon^{-kx} \cdot \epsilon^{-\gamma x} \cdot dx. \tag{5}$$

the integration extending over the entire cable. A is a factor which takes account of the shielding by armor wires, and changes at each point on the cable where the armoring changes. z is the depth of immersion at a distance x from the terminal, the relation between z and x being obtainable from the profile curve of the cable route. By comparing the susceptibility-frequency curves for two cables we can obtain an idea of the relative disturbances to be expected on the cables, with the possible exception of that part arising from sources in close proximity to the cables. For the latter type of interference special considerations are necessary.

In drawing conclusions from a susceptibility-frequency curve it is essential to bear in mind that, although the disturbance at the cable terminal is a composite of sinusoidal voltages and currents of all frequencies from zero to infinity, we are principally concerned with the

[6] An interesting conclusion to be drawn from equation (4) is that the contributions from various portions of a long section of cable due to a uniform disturbance tend to neutralize each other, on account of the fact that they arrive at $x=a$ in various phases. Since γ is equal to $\alpha+j\beta$, where α is the attenuation constant and β the phase constant, both per unit length, the quantity $\epsilon^{-s\gamma}$ can be represented graphically by a vector of length $\epsilon^{-s\alpha}$ and angle $(-s\beta)$. If α were zero the value of the factor $1-\epsilon^{-s\gamma}$ would be zero for $s\beta=o$, 2π, 4π, 6π, etc. That is the disturbance picked up over a length of cable $s=\frac{2n}{\beta}$, n being any integer, would have no effect at the terminal of the cable. On account of the fact that α is not zero, the quantity $\epsilon^{-s\gamma}$ is less than unity for all the above values of s except $s=o$, and complete neutralization of the disturbance does not occur. In the case of an inductively loaded cable, however, for a given value of α, β is many times greater than the value for the corresponding non-loaded cable. This means that neutralization of interference picked up on the loaded cable is much more complete than in the case of a non-loaded cable.

components lying within a certain frequency range, the limits of which depend upon the speed of signalling. This is due to the fact that the characteristics of an ordinary submarine cable are such that the low frequency components of a signal are transmitted with much less diminution of amplitude than are the higher frequency components. Consequently [7] it is found necessary, in order to render the signal intelligible, to employ a correcting network at the receiving terminal, one function of which is to attenuate the arriving low frequency components so that they finally are in the proper proportion to the higher frequency components. Also it is found that frequencies which are higher than about one and one-half times the signal frequency are not required in order to obtain intelligible signals, so that the receiving network can be designed to remove disturbances of the higher frequencies. The receiving apparatus therefore acts as a band filter towards the interference arriving at the terminal and emphasizes the part played by the components of interference of frequencies in the neighborhood of the signal frequency. On this account it is possible, in the majority of cases, to obtain the significant portion of the susceptibility-frequency curve by limiting the integration in (5) to the portion of the cable submerged to a depth of approximately 1000 feet or less, since, as has been previously indicated, only disturbances of extremely low frequencies are picked up on the deep water portion of the cable.

Given the problem of predetermining the interference at the terminal of a projected cable, the following procedure can be employed:

1. Over a period of time sufficiently long, a series of records of interference is taken on a cable terminating in the same general neighborhood as the proposed cable. Oscillographic records of the type shown in Fig. 1 are very desirable for this purpose.

2. From these records, and from the computed susceptibility-frequency curves of the existing and projected cables the interference on the latter can be predicted.

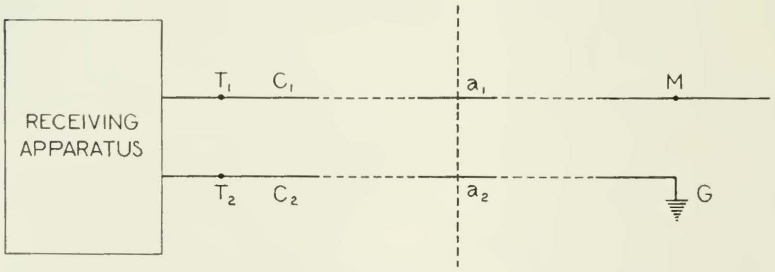
The method just described was applied to predetermine the interference at the terminals of the New York-Azores permalloy loaded cable. At the Azores terminal the cable reaches deep water within a few miles of the terminal, and the results indicated that the magnitude of interference to be expected would be sufficiently small to permit of signalling at the speed at which it was desired to operate. At the New York terminal, however, the ocean for a distance of about 100 nautical miles is comparatively shallow, and cables in this vicinity are exposed to rather severe disturbances. This is partly due to unusually strong

[7] See Milnor "Submarine Cable Telegraphy," Trans. *A. I. E. E.*, Vol. 41, p. 20, 1922.

stray fields from the numerous electric railway systems in the neighborhood of New York. By means of an amplifier and a recording string oscillograph records were obtained of the interference on the Western Union Telegraph Company's non-loaded cables terminating at New York. In taking these records a number of terminal networks were employed, with various attenuation characteristics, in order to obtain an idea of the distribution of interference with respect to frequency. Another series of tests was made, on board the Western Union cableship "Clowry," during which a cable was raised from deep water, cut, and interference studies made on the two parts of the cable. A study of these results according to the method that has just been described indicated that unless some means were employed for reducing the terminal interference, a great sacrifice of signal speed would have to be made, at least on westbound traffic. The remedy that was adopted is a special type of earth connection 100 miles at sea, to which the ground terminal of the receiving apparatus is connected. The theory of this arrangement will now be developed.

For the purpose of diminishing extraneous interference there is provided on most submarine cables an earthing arrangement, which, as shown diagrammatically in Fig. 5, consists of a core C_2 of the same general type as that used in the main cable C_1, and which may be

FIG 5



armored either with the main cable or in an independent sheath. This core usually extends for a distance of a few miles from the shore, to a point G, where the conductor of the core is grounded on the armor of the main cable. The receiving apparatus associated with the main cable conductor is then connected to earth through the sea earth conductor and the earth connection at its sea terminal. It is evident that if the main core and the sea earth core are close together they will both be exposed to the disturbances encountered between the terminal and the point where the sea earth conductor is grounded. If the two cores

reacted in the same degree to these disturbances, then it is clear that corresponding to each disturbing impulse at T_1 due to pick-up at any point a_1 on T_1M there would be an equal impulse at T_2 due to pick-up at a_2 the corresponding point on $T_2\,G$ and there would be no resulting difference of potential impressed on the receiving network due to these disturbances. As a matter of fact the section $T_2\,G$ does not react to disturbances in the same manner as the section $T_1\,M$, even though the two cores have identical linear characteristics. Although the impedances looking landward from a_1 and a_2 will be equal, the impedances looking seaward from the two points are likely to be widely different, and the impedances into which electromotive forces induced at a_1 and a_2 work will not be equal. The same disturbance will therefore set up currents of different amplitudes in the two conductors, and there will be a difference of potential between T_1 and T_2 which will be indicated on the receiving instrument. Another way of looking at this effect is to consider the disturbances picked up at a_1 and a_2 as resulting in transient waves of potential and current which are propagated along the two conductors in both directions from the points of pick-up. The waves travelling from a_1 to T_1 are equal to the corresponding waves travelling from a_2 to T_2. A similar equality holds for the waves travelling from a_2 to G and from a_1 to M. On arriving at G the waves on the sea earth conductor are reflected and travel back along the conductor, finally arriving at T_2. Since there is no corresponding reflection on the main conductor, there will be an unbalanced disturbance, the magnitude of which depends upon the amount by which the disturbance was attenuated in travelling over the route a_2-G-T_2.

The remedy [8] for the condition just described is to eliminate reflection at the sea end of the sea earth conductor, or, if for any reason, there is a reflection at the point M, to balance it with an equal reflection at the point G. This can be done by grounding the sea earth conductor at G through a network having an impedance that bears the same relation to the impedance of the conductor $G\,T_2$ as the impedance of the cable seaward of M bears to the impedance of the conductor $M\,T_1$. When the two cores $T_1\,M$ and $T_2\,G$ are alike, the impedance of the network should equal the impedance of the main cable at M.[9]

[8] Osborne, U. S. Patent 1,390,580—1921.
Heurtley, Br. Patent 198,978—1923.
Gilbert, Br. Patent, 218,261—1926.

[9] There is one important type of disturbance which has not been dealt with in the preceding discussion, namely, that due to the signal currents on cables which cross or lie close to the cable in which we are interested. It is evident that the electromotive forces induced in the cable conductor due to such causes behave in the same manner as any other disturbing electromotive force and that the magnitude of their effect can be reduced by the use of a balanced sea earth conductor terminated at a point beyond the region of disturbance.

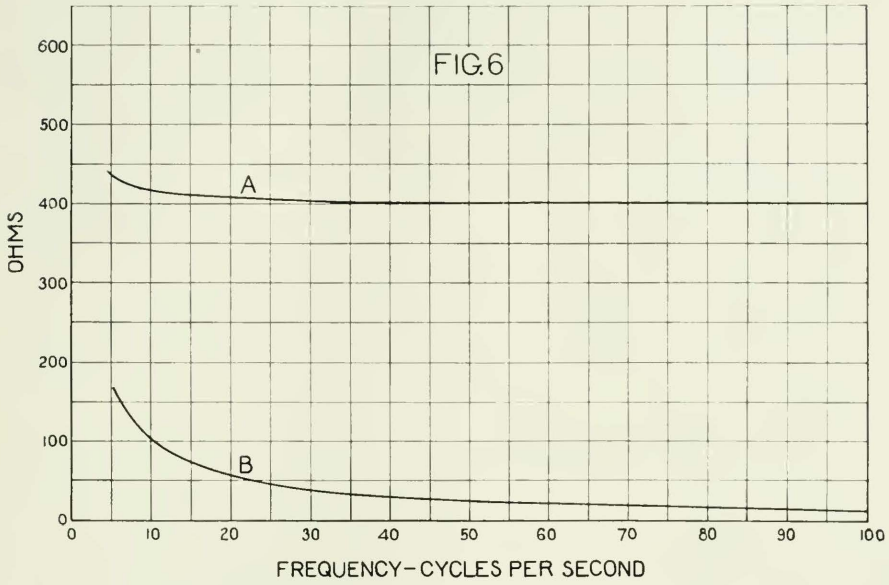
With this source of unbalance between the main core and the sea earth core removed or greatly reduced, it becomes increasingly important that the factors affecting the pick-up and the transmission of interference on the two cores be made as nearly as possible the same. In manufacturing the cable, core lengths should be paired off in such a manner that the electrical constants of any portion of the sea earth core match the constants of the corresponding portion of the main core, and the two cores should preferably be armored together.

By an extension of the method employed in deriving formula (5) an expression for the interference-susceptibility frequency characteristic of a cable having a balanced type of sea earth can be derived. This expression will consist of two terms, the first representing the resultant interference due to lack of balance between the sea earth conductor and the main core, and a second term, similar in form to (5), representing the interference picked up on the portion of the cable beyond the sea earth termination. Because of the difficulties involved in balancing, there is a value below which the first term cannot practically be reduced, which residue amounts to a few per cent of the magnitude of interference that would be encountered on this portion of the cable if the balanced type of sea earth were not employed. The second term can be reduced to any desired value by terminating the sea earth in water of sufficient depth. It is evident that when the sea earth has been extended to a point where the second term is small compared with the first, the limit of interference reduction is reached.

The question as to how far from shore the sea earth should be located in a particular case is an economic problem, the optimum location being that where the increase in value of the cable, due to diminution of interference by further extension of the sea earth, balances the additional cost of making the extension. In some cases it is found economical to obtain the desired ratio of signal-to-interference by means of a more efficient and expensive core rather than by an extended sea earth conductor. In the case of transatlantic cables terminated at points on the English Channel, or on the North Sea, for example, sea earth conductors several hundred miles in length are required in order to get a deep water termination. By increasing the weight of the main conductor, thereby increasing the amplitude of signals received over the cable, a greater amount of interference can be tolerated, in which case a comparatively short sea earth can be employed, just long enough to get rid of local interference and of the pick-up of signals from cables terminating nearby.

An inductively loaded submarine telegraph cable possesses characteristics which make the balanced type of sea earth particularly

adaptable. Fig. 6 shows the real and imaginary parts of the characteristic impedance of a typical cable designed to operate at a speed corresponding to about 60 c.p.s. It is evident that for all frequencies above 20 c.p.s. the impedance can be approximated very closely by a pure resistance of about 400 ohms. In contrast to this, the character-



A. Resistance Component. *B*. Reactance Component (negative)

istic impedance of a non-loaded type of cable varies with frequency and has a reactance component about as large as the resistance component. In the case of the loaded cable the problem of designing a terminating network for the sea earth conductor is therefore comparatively simple, being a matter of finding a method of including in the cable structure a resistance of several hundred ohms. It is true that a network of this sort does not provide a good balance for frequencies much below 20 c.p.s., and components of interference of these low frequencies will be found at the cable terminals due to the lack of balance between the main cable and the sea earth. As was previously pointed out, however, these components will be so greatly attenuated by the signal correcting networks that their effect upon the receiving instrument will be inappreciable. This is illustrative of a general property of the loaded telegraph cable, namely, that when a cable is suitably designed for the frequency at which it is to be operated its characteristic impedance approximates closely to a resistance over a

range of frequencies which extends considerably below the signal frequency, so that the resultant interference due to employing a resistance termination for the sea earth conductor will be attenuated to such a degree by the signal correction networks that it will in general have a negligible effect upon the receiving instrument. Moreover, it is probable that a considerable amount of low frequency disturbance is picked up beyond the sea earth and the gain obtained by improving the balance for these frequencies would not be very great.

A practical design for the terminating resistance consists of a length of several hundred feet of stranded wire, approximately 0.05 inch in diameter, of high resistivity material, insulated with gutta percha. After being joined at one end to the sea end of the sea earth core, the insulated conductor is served with jute and laid up with the main core for armoring exactly in the same manner as any other portion of the sea earth core. The free end of the conductor is grounded by connecting to the armor wires in the usual manner. A structure of this sort satisfies very completely the requirement of simplicity and lightness, and is as easily maintained as a length of ordinary cable similarly located.

There is a second characteristic of the loaded type of cable that tends to simplify the problem of the design of a balanced type of sea earth. It has been shown that the portion of the extraneous interference that it is most desirable to eliminate consists of the components of

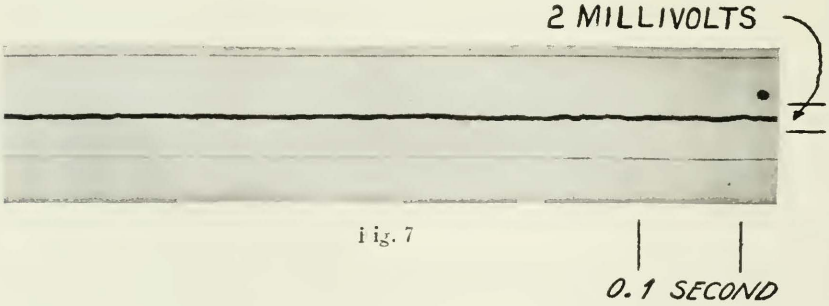


Fig. 7

frequencies in the neighborhood of the signal frequency. Since the operating speed of a loaded cable is five to ten times that of the corresponding non-loaded cable, it is evident from the preceding discussion that in order to effect a given reduction of interference in any particular locality, the sea earth of the loaded cable can be located closer to shore and in shallower water than in the case of the non-loaded cable.

In the case of the New York-Azores cable the balanced type of sea earth has been very effective in reducing extraneous interference. Fig. 7 is an oscillographic record of the terminal interference between

this cable and its sea earth taken at the same time and under the same conditions as Fig. 1, which is the record of terminal interference on an adjacent non-loaded cable provided with the ordinary type of sea earth. In both cases a large condenser was inserted between the cable and the amplifier to reduce the "zero wander" due to components of very low frequency. Comparison of the two records indicates that the interference on the cable with the ordinary sea earth is about ten times that on the cable with the balanced sea earth. The contrast between the two types of sea earth is still more pronounced at times when terminal interference is unusually large. It has been found possible, for example, to operate the New York-Azores cable during violent local electrical storms when neighboring cables were compelled to cease operation.

Neutralization of Telegraph Crossfire

By R. B. SHANCK

SYNOPSIS: With the simple means here described for neutralizing mutual interference between parallel telegraph circuits, it has been found practicable to effect a reduction to 10 or 20 per cent of the original values. This has improved considerably the operation of some circuits and made available others which were previously unsuitable. The resulting improvement in transmission has made possible the elimination of certain intermediate telegraph repeaters with material savings. The neutralizing apparatus has no material effect when crossfire is not present, that is, when the paralleling wires are idle. It has been found that the use of arrangements here described on certain long open wire circuits makes possible fast manual full-duplex operation where only medium-speed half-duplex operation was possible before. Furthermore, in the case of some cable circuits where it was impossible to operate more than two telegraph circuits per quad, it is now practicable to obtain four telegraph circuits.

INTRODUCTORY

MANY ground-return telegraph circuits are subject to serious mutual interference due to their proximity to one another on pole lines or in cable and in certain cases due to interconnection in office apparatus. The interfering currents, commonly referred to as "crossfire", in one telegraph circuit, caused by the transmission of signals on paralleling telegraph circuits, have caused considerable difficulty in the operation of such circuits. Crossfire has either limited the speed of operation or seriously impaired the quality of transmission in many cases.

In the following there are described methods which have been successfully applied to a number of ground-return polar-duplex telegraph circuits in the Bell System for the purpose of neutralizing crossfire. These arrangements are comparatively inexpensive and afford a marked improvement in transmission. This paper deals specifically with methods for use on wires which are either used simultaneously for telephone purposes or at least are grouped and transposed so as to be suitable for telephone operation; there is, however, no reason why the principles may not be profitably applied in many cases where wires are intended exclusively for telegraph use.

NATURE OF CROSSFIRE

When mutual admittance, or coupling, exists between two telegraph circuits, operation of one, of course, occasions extraneous current impulses in the other circuit. The presence of such impulses in the receiving apparatus at the terminals of the disturbed circuit results in adverse effects on the telegraph signals. In the case of

closely parallel circuits extending between two stations, considerable interference is generally experienced both at the station from which the disturbing signal is transmitted and also at the distant station. In this paper, the crossfire current (noted in the interfered—with circuit) at the station from which the interfering signal is sent will be referred to as "sending-end crossfire" and that at the distant station as "receiving-end crossfire". For example, assume two parallel wires from A to B; if a signal be sent on wire No. 1 from A to B, sending-end and receiving-end crossfire will appear in the receiving apparatus of wire No. 2 at A and B, respectively. This may mutilate incoming signals, or in extreme cases cause false signals.

The type of line circuit and the kind of apparatus employed have a considerable effect upon the amount of crossfire between circuits. It has been found that it depends chiefly upon the amount of mutual capacitance and, to a lesser extent, upon the natural mutual inductance of the wires; mutual conductance or leakage is responsible for some d-c. crossfire during periods of low insulation resistance but this increment is in general comparatively unimportant. As will be brought out later, loading[1] of circuits has a large effect on crossfire. Such factors as the gauge of wire, separation between wires, length of circuit and the presence of other wires on the same pole line have, of course, considerable influence.

In the Bell System plant, crossfire is in general of little consequence except among the four wires of a "phantom" group, the reasons for which will be discussed later. It is of interest to note that receiving-end crossfire is comparatively much more serious between wires in cable than between those of open-wire lines. Entrance cable, that is, cable employed to bring open-wire circuits into large cities, has comparatively little effect, as the length is generally short. Such apparatus as the composite sets which are used to derive d-c. telegraph circuits from telephone wires, and in certain cases filters used in connection with superposed carrier-current systems, contribute to crossfire inasmuch as they introduce some coupling, chiefly mutual capacitance.

Fig. 1 shows schematically the circuit arrangement of the polar-duplex telegraph apparatus in conjunction with a pair of wires composited for simultaneous telephone and telegraph operation. These types of apparatus are well known and will therefore be described only briefly. Independent two-way telegraph transmission is possible on each wire since the receiving relay occupies a position in a

[1] See "Development and Application of Loading for Telephone Circuits", Shaw and Fondiller, A. I. E. E. Jour. March, 1926.

balanced circuit analogous to that of a Wheatstone-bridge galvanometer as regards outgoing signals, and is therefore operated only by incoming signals. The fact that the home relay is not influenced by the home battery will be appreciated by considering the battery at the distant station to be short-circuited and the artificial line to balance the line and distant apparatus perfectly. When battery is introduced at the distant station however it causes a signal to be

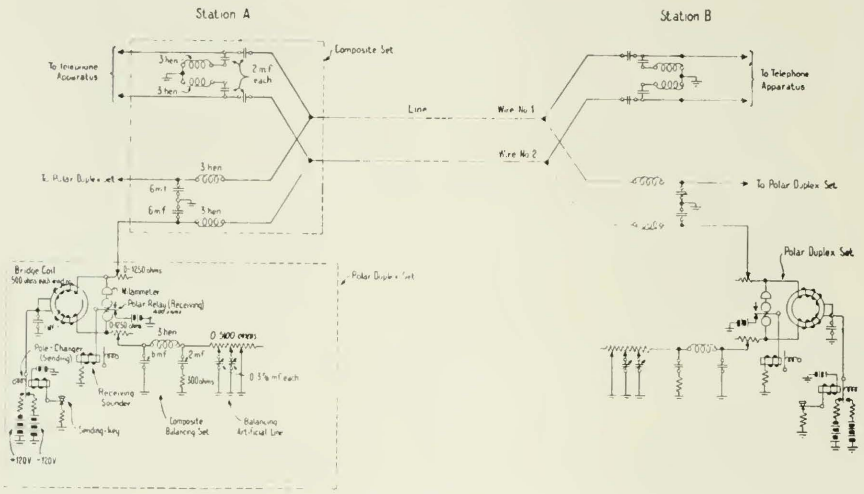


Fig. 1—Line equipped with composite sets and ground-return polar duplex sets

received at the home station as it corresponds to inserting a battery in one arm of a Wheatstone bridge. The "bridge coil" is connected so as to be series-aiding for incoming signals, the inductance being about 75 henries in this case, and parallel-opposed for outgoing signals, the inductance then being about 3 henries. The composite set serves to separate the telephone currents from those of the two telegraph circuits by "filtering" action or frequency discrimination, the telegraph employing a frequency range below that of the telephone.

The oscillograms which are shown in Fig. 2 illustrate the wave shape and magnitude of crossfire impulses in comparison with the normal operating currents, in a typical composited large gauge cable circuit. Trace A shows the wave-shape of the normal current in the line at the sending end (reduced to about one-seventh as compared to the other waves), and B shows the current in the receiving relay at the distant station. Although it is not outstanding in the

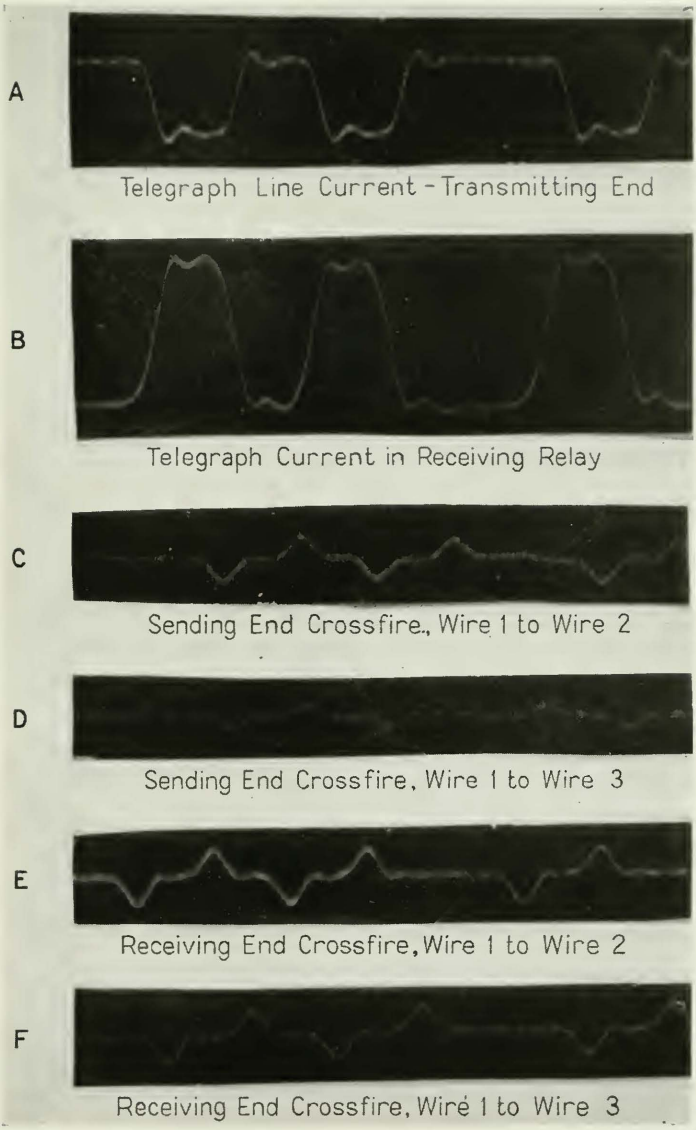


Fig. 2—Telegraph operating and crossfire currents. 13 B.&S. Ga. loaded cable Quad. 90 miles in length

Note 1: Oscillograms of crossfire current taken with vibrator in series with receiving relay

Note 2: Wave "A," 150 milliamperes per inch; other waves 20 milliamperes per inch

present instance, the sending-end current is usually characterized by peaks and rapid changes, while the received wave is somewhat rounded off, and this results in most of the induction taking place in the portion of the line near the sending station and the apparatus at that station. C illustrates sending-end crossfire between wires of the same pair and D that between wires of different pairs but in the same quad. Trace E shows the receiving-end crossfire between wires of the same pair and F that between wires of different pairs but in the same quad. C, D, E, and F may be considered as superposed in various combinations on B to obtain an idea of the mutilation of signal waves at usual speeds of manual Morse operation.

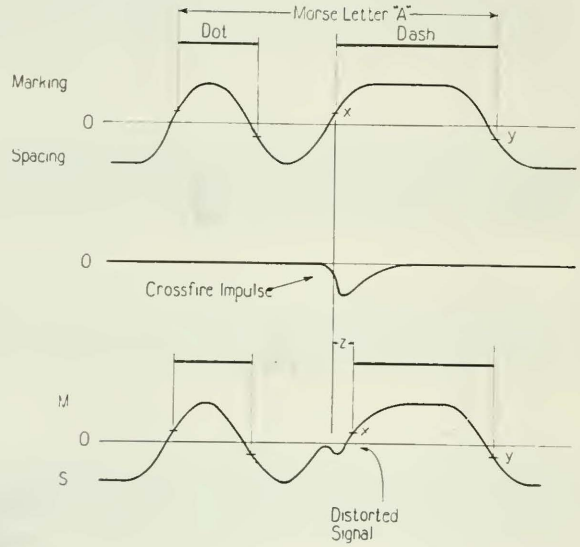


Fig. 3—Distorting effect of crossfire impulse

Fig. 3 has been drawn to illustrate how a crossfire impulse may cause distortion of a telegraph signal. The lowest wave is a combination of the received signal and crossfire impulse which are shown above. X and Y are the points at which the polar relay operates, assuming that it is required that the current build up or down appreciably beyond zero in order to move the armature. It will be clear that the dash has been shortened by the amount Z. Obviously only a limited amount of such distortion is allowable in telegraph signals. Under some conditions the crossfire is of sufficient strength to cause false signals, such as an extraneous dot in a long space or a break (space) in a dash.

An additional serious effect of crossfire is that it interferes with the obtaining of accurate duplex balance adjustment, since crossfire currents mask the effect of small changes in the balancing artificial line.

Principles of Neutralizing Arrangements

The principles involved in neutralizing the crossfire will first be discussed for the simplest case, that is, with only two parallel wires, reserving the case of four wires for the next section of this paper.

Sending-end Crossfire

An arrangement suitable for neutralizing the sending-end crossfire between two polar-duplex circuits is illustrated in Fig. 4. The

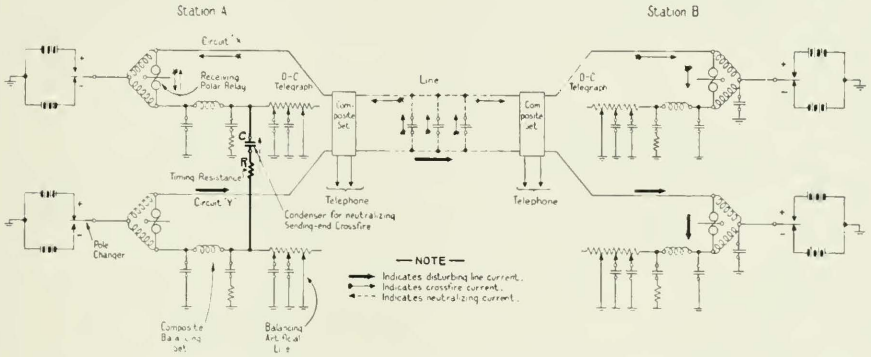


Fig. 4—Method of neutralizing sending-end crossfire between two telegraph circuits

heavy arrows indicate the disturbing line current which flows when the tongue of the pole-changer of circuit Y at station A moves from the negative to the positive pole. The feathered arrows show the direction of the resulting crossfire currents which tend to flow through the polar relays of circuit X. It will be apparent that the sudden increase, in a positive direction, of the potential applied to circuit Y would cause an impulse of current in the relay of X at the sending station A in the direction shown if the circuits were coupled by capacity only or by the natural mutual inductance of the two parallel ground-return circuits. Neutralization is effected by providing a mutual admittance between the two balancing artificial lines to simulate that existing between the real lines. It will be clear that upon the operation of the pole-changer of circuit Y, an impulse will pass through the neutralizing circuit C, R, and through the relay of circuit X at A in such direction as to oppose the crossfire current.

(The neutralizing impulse is indicated by the dotted arrows.) Another point of view is that a symmetrical or balanced arrangement similar to a Wheatstone bridge is provided in which the coupling of the line circuits is balanced by the coupling introduced between the artificial lines. It has been found experimentally that a simple connection consisting of a condenser and a timing resistance in series as shown are sufficient to effect neutralization on either open-wire or cable circuits. It will, of course, be seen that such a connection is effective for neutralizing crossfire from either circuit into the other, and furthermore that it is capable of performing both functions simultaneously.

As shown in Fig. 4 the neutralizing connection is made at the beginning of the artificial line (at the junction of it and the composite balancing set). This is a convenient point and has been found satisfactory for the purpose.

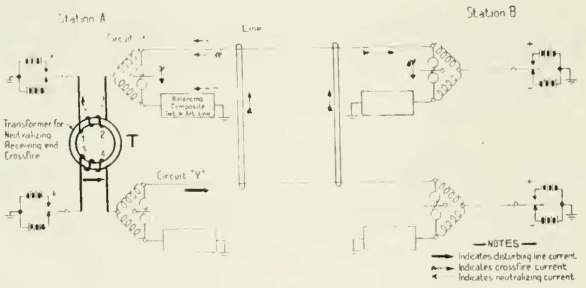


Fig. 5—Method of neutralizing receiving-end crossfire between two telegraph circuits

Condenser arrangements have been in use in this country and abroad for some years in various ways for neutralizing sending-end crossfire on both land lines and short submarine cables.

Receiving-end Crossfire

For neutralizing receiving-end crossfire use is made of special connections at the sending end. The method consists in impressing a neutralizing impulse on the disturbed circuit at the sending station, in such manner as not to affect incoming signals at that station, (that is, it does not introduce sending-end crossfire); the neutralizing impulse will then travel along the interfered-with circuit so as to arrive at the distant station at the time when the crossfire impulse appears at that station.

The operation of the receiving-end crossfire neutralizing apparatus will be made clear by reference to Fig. 5, in which the heavy arrows

indicate the disturbing current, the feathered arrows the crossfire current and the dotted arrows the neutralizing current. The last mentioned current is impressed upon the disturbed circuit X by means of a transformer connection (T) between the "apex" or transmitter branches of the two circuits. It will be obvious that if a good duplex balance has been obtained the neutralizing impulse will divide practically equally between the real and artificial lines of circuit X and substantially none of it will pass through the receiving polar relay of circuit X, on account of the balanced bridge arrangement. It will therefore have no effect on signals received at

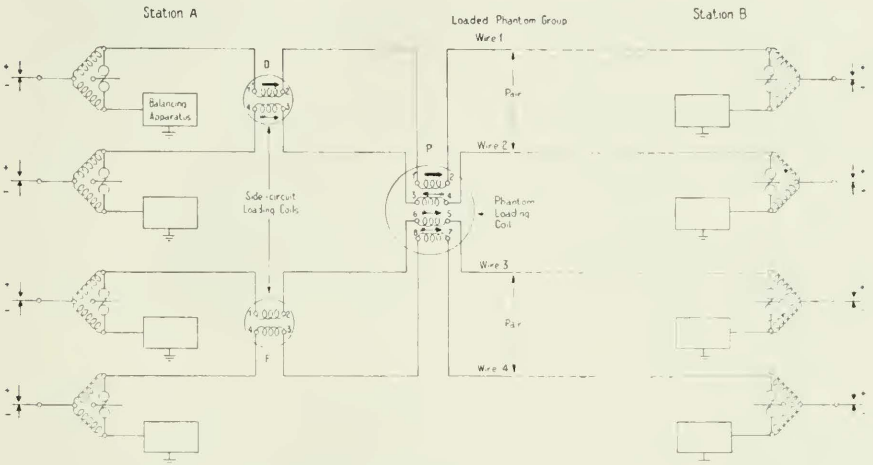


Fig. 6—Effect of loading coils on crossfire

Note: → Signalling Current ➡→ Induced Current

A, but will generate neutralizing impulses which will travel over X to B so as to appear at B at the same time as the crossfire currents. It has been found possible to employ coupling such that the receiving-end crossfire is practically eliminated in the polar relay. The proper poling of the neutralizing transformer has been found to be as indicated in Fig. 5 for all types of circuit to which the device has been applied. The crossfire impulse has the direction shown, for the reason that capacity coupling predominates.

Line Characteristics

The first part of this section will be devoted to a discussion of the effect of loading and line transpositions. This will show why, in the telephone plant, it is necessary to deal with crossfire among

the four wires of a phantom group only. Then arrangements for use with a group of four wires employing the principles explained above in connection with the case of two parallel wires, will be covered.

It is of interest to consider the effect of the loading coils which are employed in conjunction with many telephone lines.[2] In Fig. 6, coils D and F represent side-circuit loading coils on pairs 1–2 and 3–4, respectively, and P a phantom-circuit loading coil. Such coils are connected into telephone circuits at intervals to introduce inductance into the two telephone side circuits and the phantom telephone circuit, respectively.

The action of the side-circuit coil, (D), will first be considered. If a positive telegraph impulse is sent from A to B over wire 1, as indicated by the heavy arrow, it is evident that the coil acting as a transformer will set up a crossfire current in wire 2 in the same geographical direction, as indicated by the feathered arrow. The relation of this impulse to those due to capacity coupling is of interest since the capacity effect predominates. Comparison with Fig. 5, will show that at the transmitting station the impulse due to coil D will oppose the sending-end crossfire which is due to capacity coupling between circuits, while at the distant end it will augment the receiving-end crossfire due to capacity coupling. Coil F functions similarly in pair 3–4.

In the case of the phantom loading coil (P) sending an impulse from A to B on wire 1 results in disturbing currents in the same geographical direction in wires 3 and 4 and in the opposite direction in wire 2, since the coil is connected so that two windings are series-opposed in each side circuit and parallel-aiding in the phantom circuit. Comparing with Fig. 5, as before, it will be seen that for wires of a group but not of the same pair these coils tend to neutralize the sending-end crossfire which is due to mutual capacitance and augment the receiving-end crossfire due to capacity coupling; the conditions will be reversed however for wires of the same pair.

In the case of loaded circuits, crossfire, therefore, is due to loading as well as to the mutual capacitance and inductance of the wires and the coupling which exists in office apparatus, so that the final result is difficult to predict. Work with loaded circuits, which has been largely confined to cables indicates that on such circuits receiving-end crossfire is generally greater than sending-end crossfire, and sending-end crossfire between wires which are in the same phantom group but not in the same pair is so small as to be almost negligible.

Line transposition of telephone circuits has been discussed at

[2] Shaw and Fondiller, loc. cit.

considerable length in a previous paper.[3] Such transpositions consist in interchanging systematically the pin positions of the two wires of a pair and of the wires of the two pairs comprising a phantom circuit. It should be clearly understood that while these transpositions are effective in balancing a two-wire or metallic circuit against other circuits, they cannot be used to balance ground-return circuits (such as the telegraph circuits in question) against each other; however, their effect in varying the separation of the different wires from each other has a great influence on the coupling between the ground-return circuits.

A possible transposition section for an open-wire phantom group is shown in Fig. 7. It will be seen that the two wires of a pair are

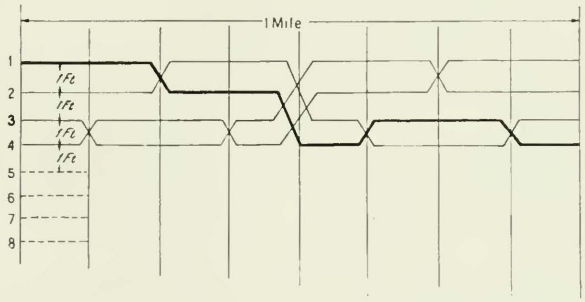


Fig. 7—Line transpositions of open-wire phantom group

always adjacent to each other and will therefore have considerable coupling; a wire of one pair is adjacent to a particular wire of the other pair for only one-fourth of the distance, and wires of the two pairs will therefore have much less coupling.

A brief consideration will make it clear that coupling between wires of separate phantom groups is comparatively small. Each wire of the group 1 to 4 occupies pin position 4 only one-fourth of the distance, and if 5 to 8 be phantomed each wire of the latter group will use pin position 5 one-fourth of the distance. It follows that a wire of group 1 to 4 will be adjacent to a particular wire of group 5 to 8 only one-sixteenth of the distance in a long circuit. If 5-6 be non-phantomed however each wire of the pair will use position 5 half of the time and will be adjacent to each wire of 1 to 4 one-eighth of the way. The next crossarms above and below are each two feet distant and carry wires transposed so as to minimize the coupling.

[3] "The Design of Transpositions for Parallel Power and Telephone Circuits," H. S. Osborne, Proc. A. I. E. E. 1918, Vol. XXXVII p. 739.

It should be noted that in addition to the reduction in coupling due to increasing the spacing there is a large reduction due to shielding when a third conductor is interposed between two others.

In the case of cable circuits the wires are twisted in groups of four so as to be transposed practically continuously. On account of the

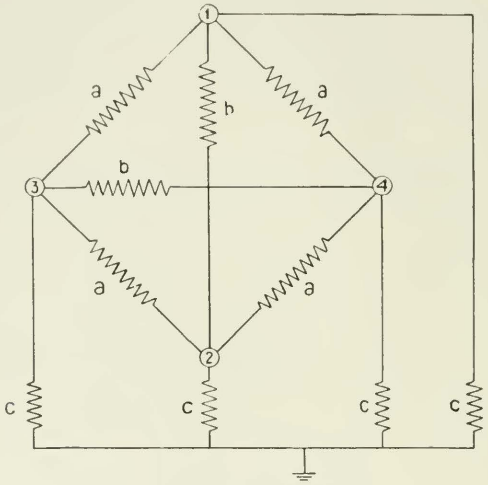


Fig. 8—Admittance Network

smaller separation, mutual capacitances and the resulting crossfire among wires of a phantom group, are considerably greater than in open wire.

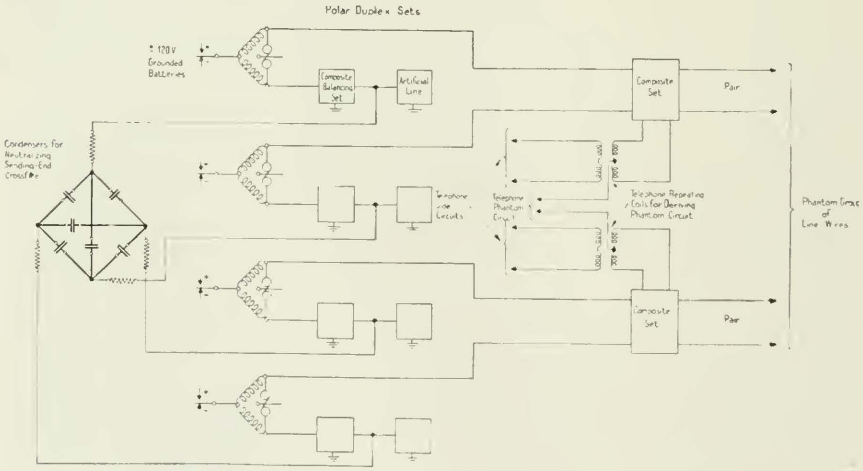


Fig. 9—Condenser arrangement for neutralizing sending-end crossfire between telegraph circuits on a phantom group

The result of transposing is that for practical purposes in connection with the crossfire problem the other wires of the line can be ignored and a phantom group represented by a network of admittances as shown in Fig. 8, where 1 and 2 represent a pair and 3 and 4 the other pair.

A network of the form of Fig. 8, is used as shown in Fig. 9, for neutralizing sending-end crossfire among the four wires of a phantom

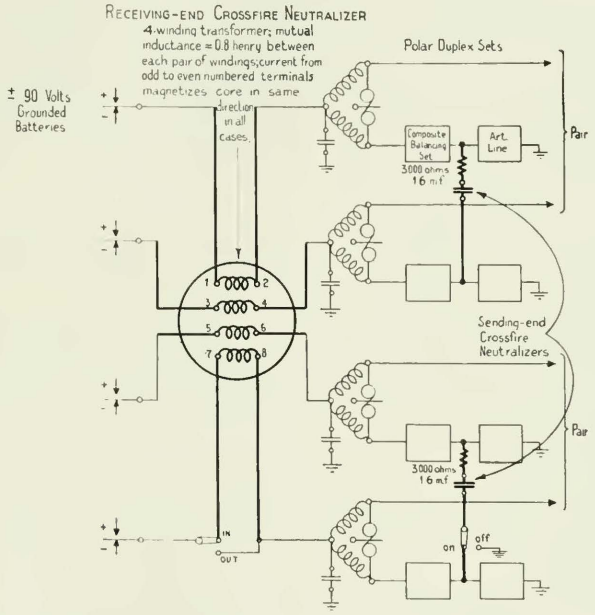


Fig. 10—Arrangement for neutralizing crossfire between telegraph circuits on a loaded No. 13 B.&S. Ga. phantom group 90 to 120 miles long in cable

group. The grounded branches shown in Fig. 8 are omitted, however, since the duplex artificial lines themselves constitute these branches. The six-mesh network consists of condensers, the timing resistances when used being external to the network as a matter of convenience.

For coupling together the apex circuits of the four wires to neutralize receiving-end crossfire, it is possible to use a special four-winding transformer analogous to the six-condenser network, but where an ordinary transformer as shown in Fig. 10 will not suffice it is convenient to employ two or three transformers in combination. For example a two-winding transformer may be added to each pair of the arrangement illustrated in Fig. 10 so as to provide additional coupling between the two wires of a pair.

It will readily be seen that with neutralizers applied at each end of the four circuits, transmission of signals on one of them will generate the proper impulses for neutralizing both sending and receiving-end crossfire from that circuit into the other three. Furthermore, neutralization will take place with all wires operating simultaneously in either or both directions.

Application to Different Circuits

It is in general not practicable to compute the constants of the neutralizing devices, but this is unnecessary since it is an easy matter to determine them experimentally. In making trials to determine the proper amount of capacity and inductance required to neutralize crossfire effectively, it is fortunately possible to design the various parts independently of each other to a considerable extent. For example, the diagonals of the six-condenser network may be determined after the condensers in the sides have been approximated very roughly, or vice-versa; likewise, the amount of inductance required between each pair of circuits may be approximated independently, but if a single coil is to be used for coupling more than two circuits, all the circuits should be connected up in making the test. Sending and receiving-end crossfire may, of course, be treated separately.

It is convenient to vary the capacity of the condensers, but not usually the inductance of the transformer. In the latter case a coil with excess mutual inductance may be used together with a variable resistance shunt.

In order to design neutralizing arrangements or to determine whether or not they are effective, tests may readily be made by observing the deflection of a milliammeter connected in series with the polar relay of a bridge polar-duplex set while signals are sent on the parallel circuit. In a similar way a differential meter may be used in a differential duplex set. A somewhat more accurate test may be made by observing the response of the receiving relay, preferably with variable electrical bias. The disturbing signals are of course sent from the same station in checking sending-end crossfire and from the distant station in checking receiving-end crossfire.

Representative anti-crossfire capacity values are given in the following table for No. 8 B.W.G. (0.165 in., 2.5 mm.) composited open-wire copper circuits, 300 to 500 miles (500 to 800 km.) in length and No. 12 A.W.G. (0.104 in., 1.5 mm.) circuits 150 to 300 miles (250-550 km.) in length. No timing resistance is required usually. In practice there are material variations from one circuit to another.

Gauge	Loading	Non-Phantomed Pair	Phantom Group Network[4] Diagonals	Sides
8	Non-loaded	1.7 mf.	1.7 mf.	1.2 mf.
8	Loaded	1.1 "	1.1 "	0.55 "
12	Non-loaded	1.1 "	1.1 "	0.8 "
12	Loaded	0.8 "	0.8 "	0.4 "

The superposition of carrier-current channels by means of filters connected on the drop side of the d-c. composite set of course has no appreciable effect on crossfire. However, the use of "transfer filters" at intermediate points to transfer the carrier from one pair to another increases the coupling between wires of a pair and this may be taken care of by increasing the capacity of the diagonals of the condenser network.

The arrangement shown in Fig. 10 has been found to be suitable for use with 90 to 120 mile (145 to 190 km.) sections of No. 13 B.&S. gauge (0.072 in., 1.8 mm.) loaded cable circuits.

In the case of open-wire circuits, receiving-end crossfire is commonly not serious excepting in special cases where high-frequency carrier telephone or telegraph transfer filters are employed. In such cases, effective neutralization may be secured by coupling the wires of each pair by means of a transformer, no such coupling being provided between the wires of separate pairs of a phantom group.

In some cases the neutralizing arrangements and the telegraph repeaters have been wired to jacks in such a manner that it is possible to patch the neutralizers from set to set by means of cords when the line assignment is changed temporarily. In some cases it is not desirable to provide such elaborate arrangements and, therefore, switches are provided for disconnecting the neutralizing apparatus from each set independently. In the case of the condensers, the duplex balance of the other telegraph sets associated with the particular group of condensers is preserved by switching directly to ground the connection from the artificial line of the set to be disconnected, as shown in conjunction with the lowermost duplex set in Fig. 10. Switches for disconnecting the neutralizing transformers are illustrated also for the same duplex set in this figure.

Practical Results Obtained with Neutralization

The following table gives data which show roughly the amount of crossfire between wires of a phantom group without neutralizing

[4] See Figure 9.

arrangements for various circuit conditions. It will be noted that crossfire between a pair of wires used for a telephone side circuit is considerably greater than that between wires of a phantom group but not of the same pair. This is in accord with what was brought out above regarding coupling. The receiving-end crossfire is much greater between cable circuits than in the case of open wires, due to the greater mutual capacitance and heavier loading.

Crossfire Current in Per Cent. of Operating Direct-Current

For Average Repeater Sections

Type of Circuit	Sending End		Receiving End	
	From Other Wire of Pair	From Wire Of Other Pair	From Other Wire of Pair	From Wire Of Other Pair
Non-loaded Open Wire	20	10	10	5
Loaded Open Wire	10	5	5	5
13 B.&S.Ga. Loaded Cable	20	5	30	25

It is practicable to reduce the crossfire to 10 or 20 per cent. of the original value by means of the arrangements which have been described. This has improved considerably the operation of some circuits and made available others which were unsuitable for use. By improving transmission so as to avoid the use of intermediate telegraph repeaters material savings have been effected in certain cases.

The neutralizing apparatus has no material effect on the quality of telegraph transmission obtained when crossfire is not present, that is, with the parallel wires idle; the application of them, however, reduces greatly the detrimental effect of crossfire on transmission. For example, the use of these arrangements on certain long open-wire circuits makes possible fast manual full-duplex (two-way) operation where only medium-speed half-duplex (one-way) operation was possible before. Furthermore, in the case of some cable circuits where it was previously impossible to operate more than two telegraph circuits per group of four wires, it is now practicable to obtain four telegraph circuits per quad.

Due to reduction of crossfire, it is usually possible to secure a much better duplex balance after the neutralizers have been applied. The application of anti-crossfire condensers however requires that a somewhat different setting of the duplex artificial line be obtained for the best balance, since the extra connection has appreciable admittance to ground.

Operation of Thermionic Vacuum Tube Circuits

By F. B. LLEWELLYN

Synopsis: Given the static characteristic of grid current-grid potential, and plate current-plate potential, for any three element vacuum tube, the general exact equations for the output current when the tube is connected in circuits of any impedance whatsoever, and excited by any variable voltage, are here derived. The method of derivation is illustrated in the special case where resistances only are considered, and the adaptation of complex impedance to use in non-linear equations is shown. Approximations that are allowable in various practical applications are indicated, and the equations are applied in some detail to grid-leak detectors, and in brief to other types of detectors, modulators, amplifiers and oscillators.

Certain repetitions of previous work are contained in these pages, as it is believed that the applications of the novel features introduced are illustrated thereby better than by a description dealing only with new material.

THE equations in use at the present time for the relation between input voltage and output current in thermionic vacuum tubes are those developed by a number of pioneers in Radio Communication. They have been summarized very concisely, and somewhat extended in an important paper by John R. Carson, entitled "A Theoretical Study of the Three Element Vacuum Tube," which appeared in the Proceedings of the Institute of Radio Engineers, April, 1919. For some time past the need of relations which include the effect of the variation of certain quantities, considered constant in Mr. Carson's paper, has been growing. Especially in the case of detection and modulation has this need become pronounced. Moreover, in the special case of grid leak detectors, the need for a general theoretical analysis has not, to the author's knowledge, been completely satisfied.

Purpose

It is, therefore, the purpose of the present paper to derive general exact equations for the output current from a three-element thermionic vacuum tube when it is connected in circuits of general impedance, both on the input and output sides, and to show specific methods of applying these general equations to several special cases, with emphasis on the case of the grid leak detector. It is also proposed to show that, whether used for detectors, modulators, amplifiers, or oscillators, the same fundamental theory applies. It is hoped that the theory and methods given will form a basis upon which a complete rational design of vacuum tube circuits may be built.

Theory

In the derivation of these equations, no limitations whatever should be imposed. Consider a three-element vacuum tube connected in circuits of general impedance on both input and output sides. The

grid is allowed to take convection current. The amplification factor, μ, is considered variable, and the effect of plate potential on grid current is included. Under these conditions, the total plate current of the tube can merely be said to be a function of the grid and plate potentials; and the total grid current, likewise, is some other function of the grid and plate potentials. The fundamental relations:

$$I_p = I_p(E_g, E_p) \tag{1}$$

$$I_g = I_g(E_g, E_p) \tag{2}$$

express, the operation of the device. They represent the static characteristics of the tube. It is from these two relations alone that the general theory must be built.

In order to do this, the following notation will be employed:

$$\left.\begin{aligned} I_p &= I_{po} + i_p \\ I_g &= I_{go} + i_g \\ E_p &= E_{po} + e_p \\ E_g &= E_{go} + e_g \end{aligned}\right\} \tag{3}$$

It will be recognized that the lower case letters represent variations in the normal values of the currents and voltages denoted by the zero subscripts. It should be noted, moreover, that all voltages and currents refer to the effect directly on the element of the tube, plate or grid as the case may be.

With the aid of (3), equations (1) and (2) may be written

$$i_p = P_1 e_g + P_2 e_p + \tfrac{1}{2} P_3 e_g^2 + P_4 e_g e_p + \tfrac{1}{2} P_5 e_p^2 + \ldots \tag{4}$$

$$i_g = T_1 e_g + T_2 e_p + \tfrac{1}{2} T_3 e_g^2 + T_4 e_g e_p + \tfrac{1}{2} T_5 e_p^2 + \ldots \tag{5}$$

where the P's and T's have the following significance:

$$\left.\begin{aligned} P_1 &= \frac{\partial I_{po}}{\partial E_g} \quad P_2 = \frac{\partial I_{po}}{\partial E_p} \quad P_3 = \frac{\partial^2 I_{po}}{\partial E_g^2} \quad P_4 = \frac{\partial^2 I_{po}}{\partial E_g \partial E_p} \quad P_5 = \frac{\partial^2 I_{po}}{\partial E_p^2} \\ T_1 &= \frac{\partial I_{go}}{\partial E_g} \quad T_2 = \frac{\partial I_{go}}{\partial E_p} \quad T_3 = \frac{\partial^2 I_{go}}{\partial E_g^2} \quad T_4 = \frac{\partial^2 I_{go}}{\partial E_g \partial E_p} \quad T_5 = \frac{\partial^2 I_{go}}{\partial E_p^2} \end{aligned}\right\} \tag{6}$$

Equations (4) and (5) are obtained directly from the extension of Taylor's Theorem. The P's may be written in more useful form with the aid of (1) and the well-known definitions of the amplification

factor, μ, the plate resistance, r_p, and the grid resistance, r_g. Thus, from (1)

$$\left.\begin{aligned} \mu &= \frac{\dfrac{\partial I_p}{\partial E_g}}{\dfrac{\partial I_p}{\partial E_p}} = -\frac{dE_p}{dE_g}\Bigg]_{I_p} \\ \frac{1}{r_p} &= \frac{\partial I_p}{\partial E_p} \\ \frac{1}{r_g} &= \frac{\partial I_g}{\partial E_g} \end{aligned}\right\} \text{(by definition)} \tag{7}$$

Hence

$$\left.\begin{aligned} P_1 &= \frac{\mu}{r_p} \\ P_2 &= \frac{1}{r_p} \\ P_3 &= \frac{1}{r_p}\frac{\partial \mu}{\partial E_g} + \frac{\mu}{r_p}\frac{\partial \mu}{\partial E_p} - \mu^2\frac{r_p'}{r_p^2} \\ P_4 &= \frac{1}{r_p}\frac{\partial \mu}{\partial E_p} - \mu\frac{r_p'}{r_p^2} \\ P_5 &= -\frac{r_p'}{r_p^2} \end{aligned}\right\} \tag{8}$$

where $r_p' = \dfrac{\partial r_p}{\partial E_p}$.

In similarly treating the T's, it was found convenient to introduce an entirely new symbol. This has been done with reluctance, for it is realized that considerable difficulty has been experienced in the standardization of symbols already in use. But inasmuch as the simplification of both physical interpretation and mathematical expression which results from the use of this new symbol is enormous, its addition is believed to be warranted.

This new symbol we will call the reflex factor, and will denote it by the symbol, ν. It is analogous in its effect on the grid circuit to the effect of μ on the plate circuit. Its definition is analogous to that of μ. Thus, from (2):

$$\nu = \frac{\dfrac{\partial I_g}{\partial E_g}}{\dfrac{\partial I_g}{\partial E_p}} = -\frac{dE_p}{dE_g}\Bigg]_{I_g} \tag{9}$$

Comparison of (7) and (9) shows that while μ is equal to minus the ratio of the increments of E_p and E_g necessary to maintain the plate current constant, ν is equal to minus the ratio of the increments of E_p and E_g necessary to maintain the grid current constant. On the other hand, while in the case of μ, the ratio

$$\frac{dE_p}{dE_g}\Big]_{I_p}$$

is intrinsically negative and occurs in (7) with a negative sign, making μ intrinsically a positive number; in the case of ν, the ratio

$$\frac{dE_p}{dE_g}\Big]_{I_g}$$

is usually intrinsically positive, and occurs in (9) with a negative sign; hence ν is usually intrinsically a negative number.

With the foregoing definition, the T's may be written as follows:

$$\begin{aligned}
T_1 &= \frac{1}{r_g} \\
T_2 &= \frac{1}{\nu r_g} \\
T_3 &= -\frac{r_g'}{r_g^2} \\
T_4 &= \frac{1}{r_g}\frac{\partial}{\partial E_g}\left(\frac{1}{\nu}\right) - \frac{1}{\nu}\frac{r_g'}{r_g^2} \\
T_5 &= \frac{1}{\nu r_g}\frac{1}{\partial E_g}\left(\frac{1}{\nu}\right) + \frac{1}{r_g}\frac{\partial}{\partial E_p}\left(\frac{1}{\nu}\right) - \frac{1}{\nu^2}\frac{r_g'}{r_g^2}
\end{aligned} \tag{10}$$

where $$r_g' = \frac{\partial r_g}{\partial E_g}.$$

The effective value of T_2, when taken over a cycle of sine wave form, has sometimes been called the reflex mutual conductance (L. A. Hazeltine), and has been denoted by g_n. An attempt to adapt this notation to the present purpose has not proved feasible. For reference, it may be noted in the limiting case, where the amplitude of the sine wave approaches zero:

$$g_n = \frac{1}{\nu r_g}.$$

With the relations given thus far, the problem may now be more specifically stated as follows:

It is desired to express i_p, the output current through a general

impedance in the plate circuit, as an explicit function of e, a variable voltage applied in series with a general impedance in the grid circuit.

Special Case

The following special case will make the detailed derivation, where complex quantities are considered, more intelligible.

For this special case consider a vacuum tube connected in circuits containing only pure resistances. Let the resistance in the grid circuit be denoted by Q and that in the plate circuit be denoted by Z. Fig. 1 illustrates this circuit. Let i_p and i_g be determined to satisfy the following series:

$$i_p = a_1 e_g + a_2 e_g^2 + \ldots \tag{11}$$

$$i_g = b_1 e + b_2 e^2 + \ldots \tag{12}$$

(11) and (12) are valid since (4) and (5) are formally power series. As seen from Fig. 1, e represents a variable voltage impressed in series with the resistance, Q, on the grid of the tube.

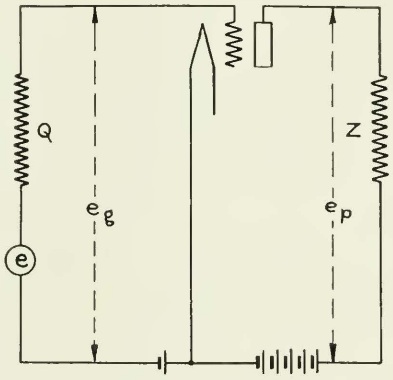


Fig. 1—Fundamental circuit diagram

These equations will give the plate and grid currents as explicit function of the voltages e_g and e, respectively, if we can evaluate the a's and b's. To do this, we have the relation

$$e_p = -i_p Z. \tag{13}$$

Substituting (11) in (4) and equating coefficients of like powers of e_g, we may evaluate a_1 and a_2 and thus express i_p as an explicit function of e_g:

$$i_p = \frac{\mu}{(r_p+Z)} e_g + \frac{1}{2}\left[\frac{-\mu^2 r_p r_p' + \mu\frac{\partial \mu}{\partial E_p}(r_p^2 - Z^2) + \frac{\partial \mu}{\partial E_g}(r_p+Z)^2}{(r_p+Z)^3}\right] e_g^2. \tag{14}$$

In equation (14), when the amplification factor, μ, is considered constant, we have the well-known relation as given in Mr. Carson's paper

$$i_p = \frac{\mu e_g}{(r_p + Z)} - \frac{1}{2}\frac{\mu^2 r_p r_p'}{(r_p+Z)^3} e_g^2 + \cdots \tag{15}$$

Experiments have shown, however, that when the resistance, Z, is not small compared to r_p the modulation resulting from variations of μ amounts to an appreciable part of the total. When the grid is maintained at a negative potential with respect to the filament, (14) may be simplified somewhat by the relation which then holds quite closely[1], namely:

$$\mu \frac{\partial \mu}{\partial E_p} = \frac{\partial \mu}{\partial E_g}.$$

Equation (14) then becomes

$$i_p = \frac{\mu}{r_p + Z} e_g - \frac{1}{2}\left[\frac{\mu^2 r_p r_p'}{(r_p+Z)^3} - \frac{2 r_p \dfrac{\partial \mu}{\partial E_g}}{(r_p+Z)^2}\right] e_g^2 + \cdots \tag{16}$$

This equation is applicable to the calculation of the output current when e_g is known, and the grid takes no convection current, as is the case in very many circuits met with in practice.

It is instructive to investigate the relative magnitudes of the two components of the second term of (16) in an actual experimental case. For convenience, the contribution of the second component of this term will be called, "μ modulation." A vacuum tube was measured and found to have the following properties under operating conditions

$$r_p = 6400$$
$$r'_p = -61.3$$
$$\mu = 5.84$$
$$\frac{\partial \mu}{\partial E_g} = .05$$

The results of applying these to (16) are shown in the following table:

Z	Total modulation	μ modulation, %
0	$.03341/10^3\ e^2$	23.35
r_p	$.00515/10^3\ e^2$	37.8
$2\ r_p$	$.00186/10^3\ e^2$	46.6
$4\ r_p$	$.000889/10^3\ e^2$	55.0

[1] See appendix I for proof of this.

This illustrates strikingly the importance of the variation of μ in modulators and detectors.

Equation (14) is expressed in terms of e_g, the voltage directly on the grid of the tube. We may derive the expression for i_p in terms of e, a voltage impressed in series with a resistance, Q, in the external grid circuit by noting that

$$e_g = e - i_g Q.$$

Hence, from (12),

$$e_g = (1 - b_1 Q)e - b_2 Q e^2 + \ldots \tag{17}$$

Therefore

$$i_p = a_1(1 - b_1 Q)e - [a_1 b_2 Q - a_2(1 - b_1 Q)^2]e^2 + \ldots \tag{18}$$

and, as in (13),

$$e_p = -i_p Z.$$

Substituting (17) and (18) into (5) and equating coefficients of like powers of e, we get

$$b_1 = \frac{T_1 - T_2 a_1 Z}{1 + T_1 Q - T_2 a_1 Z Q}, \tag{19}$$

$$b_2 = \frac{[-a_2 Z T_2 + \frac{1}{2} T_3 - a_1 Z T_4 + \frac{1}{2} a_1^2 Z^2 T_5](1 - b_1 Q)^2}{1 + T_1 Q - T_2 a_1 Z Q}. \tag{20}$$

The T's may be expressed in terms of r_g and ν with the aid of (10). The complete solution of this special case for first and second order effects is then given by (18) above, in which we have now evaluated the a's and b's.

Mathematical Digression

Before the detailed steps in the complete development of the general case, with general impedances instead of resistances, are attempted, the following digression on the use of complex quantities in non-linear equations is apposite. Included at this point, it serves a two-fold purpose; first, the notation to be used is illustrated by means of simple applications; second, it calls to mind the fundamental ideas involved in the representation of impedances by complex quantities.

Consider a current, I. If periodic, this current may be represented by a Fourier series and expressed as the sum of a number of cosine terms. Thus

$$I = I_h\left(\frac{\epsilon^{j(ht+\phi)} + \epsilon^{-j(ht+\phi)}}{2}\right) + I_k\left(\frac{\epsilon^{j(kt+\theta)} + \epsilon^{-j(kt+\theta)}}{2}\right) + \ldots \tag{21}$$

where the symbol, j, represents the imaginary, $\sqrt{-1}$. For brevity this may be written

$$I = (i_{1h} + \bar{i}_{1h}) + (i_{1k} + \bar{i}_{1k}) + \text{- - -} \tag{22}$$

where the bar over a symbol denotes the conjugate imaginary of the same symbol unbarred. If this current flows through a circuit containing resistance, self-inductance, and capacity in series, we have

$$e = RI + L\frac{dI}{dt} + \frac{1}{C}\int Idt. \tag{23}$$

Substituting for I its equivalent, as given by (21) or (22), we may write the result in abbreviated form as follows:

$$e = (z_h i_{1h} + \bar{z}_h \bar{i}_{1h}) + (z_k i_{1k} + \bar{z}_k \bar{i}_{1k}) + \text{- - -} \tag{24}$$

where

$$z_n = R + Ljn + \frac{1}{Cjn},$$

$$\bar{z}_n = R - Ljn - \frac{1}{Cjn}.$$

When the current flows through a network of impedances, we may always write the equivalent series impedance of the network. Hence equation (24) may be extended to cover the general case. It will be noted that lower case z's have been used to represent impedances in the above discussion. Throughout this paper the attempt has been made to employ the lower case letters to denote quantities which involve time, reserving the capitals for those which do not involve time. With this understanding, Z denotes a resistance, while z represents a general impedance, which, of course, varies with the time variation of a voltage impressed on it. With the aid of (24) we are in a position to treat non-linear equations by the complex method. Thus, omitting conjugates e^2 becomes,

$$\begin{aligned} e^2 = {} & z_h^2 i_{2(2h)} + z_k^2 i_{2(2k)} + 2z_h \bar{z}_h i_{2(0h)} + 2z_h z_k i_{2(h+k)} \\ & + 2z_h \bar{z}_k i_{2(h-k)} + 2z_k \bar{z}_k i_{2(0k)} + \text{- - -} \end{aligned} \tag{25}$$

which may be written

$$e^2 = e_{2(2h)} + e_{2(2k)} + e_{2(0h)} + e_{2(h+k)} + e_{2(h-k)} + e_{2(0k)} + \text{- - -} \tag{26}$$

In (25) and (26) the significance of the double subscript notation is brought out. The first symbol in the subscript refers to the order of the term, and the second refers to the frequency.

In the light of the foregoing discussion, the problem of writing the general equations for the thermionic vacuum tube may be attacked.

General Analysis

Coming back to the detailed problem in hand, we follow out the method illustrated in the special case, but must use the notation developed in the preceding section to take care of a *general* impedance, z, in the plate circuit, and a *general* impedance, q, on the grid circuit. Fig. 1 as before, shows the skeleton circuit, where, however, lower case z and q must be substituted for the capitals. Then

$$e = e_{1h} + \bar{e}_{1h} + e_{1k} + \bar{e}_{1k} + \text{- - -} + e_{1n} + \bar{e}_{1n} . \tag{27}$$

Analogous to (11) and (12):

$$\left.\begin{aligned} i_p = {} & a_{1h} e_{g1h} + \bar{a}_{1h} \bar{e}_{g1h} + a_{1k} e_{g1k} + \bar{a}_{1k} \bar{e}_{g1k} + \text{- - -} \\ & + a_{2(h-k)} e_{g2(h-k)} + a_{2(h-k)} e_{g2(h-k)} + \text{- - -} \end{aligned}\right\} \tag{28}$$

$$\begin{aligned} i_g = {} & b_{1h} e_{1h} + \bar{b}_{1h} \bar{e}_{1h} + b_{1k} e_{1k} + \bar{b}_{1k} \bar{e}_{1k} + \text{- - -} \\ & + b_{2(h-k)} e_{2(h-k)} + \bar{b}_{2(h-k)} \bar{e}_{2(h-k)} + \text{- - -} \end{aligned}$$

Hence, analogous to (13) and (17):

$$e_p = -\Sigma [a_{1n} z_{1n} e_{g1n} + \bar{a}_{1n} \bar{z}_{1n} \bar{e}_{g1n} + a_{2m} z_m e_{g2m} + \bar{a}_{2m} \bar{z}_m \bar{e}_{g2m}] \tag{29}$$

$$e_g = \Sigma [(1 - b_{1n} q_n) e_{1n} + (1 - \bar{b}_{1n} \bar{q}_n) \bar{e}_{1n} - b_{2m} q_m e_{2m} - \bar{b}_{2m} \bar{q}_m \bar{e}_{2m}] \tag{30}$$

where the summation refers to terms of different frequencies but of similar form.

From this point on, the procedure is exactly the same as that given in the special case. Coefficients of terms of like order *and frequency* are equated, and the final results are:

$$\left.\begin{aligned} i_p = {} & \Sigma a_{1h} (1 - b_{1h} q_h) e_{1h} \\ & + \Sigma [(1 - b_{1h} q_h)^2 a_{2(2h)} - a_{1(2h)} q_{(2h)} b_{2(2h)}] e_{2(2h)} \\ & + \Sigma [(1 - b_{1h} q_h)(1 - b_{1k} q_k) a_{2(h+k)} - a_{1(h+k)} q_{(h+k)} b_{2(h+k)}] e_{2(h+k)} \\ & + \Sigma [(1 - b_{1h} q_h)(1 - \bar{b}_{1k} \bar{q}_k) a_{2(h-k)} - a_{1(h-k)} q_{(h-k)} b_{2(h-k)}] e_{2(h-k)} \\ & + \Sigma [(1 - b_{1h} q_h)(1 - \bar{b}_{1h} \bar{q}_h) a_{2(0h)} - a_{1(0h)} q_{(0h)} b_{2(0h)}] e_{2(0h)} \\ & + \ldots \end{aligned}\right\} \tag{31}$$

where the summation refers to terms of different frequencies but of similar form. Note that, having $a_{2(h-k)}$ and $b_{2(h-k)}$, we may readily write the appropriate expressions for the other a_2's and b_2's by reference to the formation in equation (31).

In (31) the a's and b's are given by:

$$\left.\begin{aligned}
a_{1h} &= \frac{\mu}{r_p+z_h} \\
a_{2(h-k)} &= \frac{\frac{1}{2}\left[-\mu^2 r_p r_p' + \mu\frac{\partial\mu}{\partial E_p}(r_p^2 - z_h\bar{z}_k) + \frac{\partial\mu}{\partial E_g}(r_p+z_h)(r_p+\bar{z}_k)\right]}{(r_p+z_h)(r_p+\bar{z}_k)(r_p+z_{h-k})} \\
b_{1h} &= \frac{1-\frac{\mu}{\nu}\frac{z_h}{r_p+z_h}}{r_g+q_h\left(1-\frac{\mu}{\nu}\frac{z_h}{r_p+z_h}\right)} \\
b_{2(h-k)} &= \frac{\left\{\begin{aligned}
&\frac{1}{2}\left[-r_g r_g'\left(1-\frac{\mu}{\nu}\frac{z_h}{r_p+z_h}\right)\left(1-\frac{\mu}{\nu}\frac{\bar{z}_h}{r_p+\bar{z}_k}\right) - 2a_{2(h-k)}\frac{r_g^2}{\nu}z_{2(h-k)}\right. \\
&-\frac{\partial}{\partial E_g}\left(\frac{1}{\nu}\right)\left(\frac{\mu z_h r_g^2}{r_p+z_h} + \frac{\mu\bar{z}_k r_g^2}{r_p+\bar{z}_k} - \frac{\mu^2 r_g^2}{\nu}\frac{z_h\bar{z}_k}{(r_p+z_h)(r_p+\bar{z}_k)}\right) \\
&\left.+\frac{\partial}{\partial E_p}\left(\frac{1}{\nu}\right)\left(\frac{r_g^2\mu^2 z_h\bar{z}_k}{(r_p+z_h)(r_p+\bar{z}_k)}\right)\right]
\end{aligned}\right\}}{\left[r_g+q_h\left(1-\frac{\mu}{\nu}\frac{z_h}{r_p+z_h}\right)\right]\left[r_g+\bar{q}_k\left(1-\frac{\mu}{\nu}\frac{\bar{z}_k}{r_p+\bar{z}_k}\right)\right]\left[r_g+q_{(h-k)}\left(1-\frac{\mu}{\nu}\frac{z_{(h-k)}}{r_p+z_{(h-k)}}\right)\right]}
\end{aligned}\right\} \quad (32)$$

Discussion of General Equations

Equations (31) and (32) contain the general solution of the problem. The formulas are too long to consider all effects at one time but if we separate (31) into components and consider each component separately, useful applications may be secured.

First taking the component that gives rise to amplification effects, we get

$$\begin{aligned}
i_{p(h)} &= a_{1h}(1-b_{1h}q_h)e_{1h} \\
&= \left(\frac{\mu}{r_p+z_h}\right)\left(\frac{r_g}{r_g+q_h\left(1-\frac{\mu}{\nu}\frac{z_h}{r_p+z_h}\right)}\right)e_{1h}.
\end{aligned} \quad (33)$$

The point to be noted in this relation is that when $q_h << r_g$ we have the well-known relation

$$i_{p(h)} = \frac{\mu}{r_p+z_h}e_{1h}. \quad (34)$$

Since amplifiers are usually operated under the condition that r_g is exceedingly large, the general solution has contributed nothing new to the amplifier equations for conditions where the grid is maintained at a negative potential with respect to the filament. But for positive values of grid potential both q_h and the reflex factor, ν, enter into the calculations. It may be remarked in passing that when the grid and plate are both positive by the same amount, the absolute value of ν is approximately equal to, or somewhat less than, μ. On the other hand, when, as is usually the case, the plate potential is much greater than the positive grid potential, the magnitude of ν is much greater than μ.

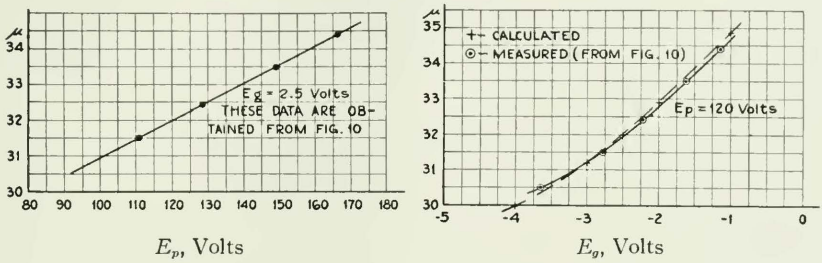


Fig. 2—Change of μ with plate and grid potentials. The points on the calculated curve were obtained as follows: since

$$\frac{\partial \mu}{\partial E_g} = \mu \frac{\partial \mu}{\partial E_p}$$

then

$$\mu = \frac{\mu_o + K E_p}{1 - K E_g}$$

where

$$K = \frac{\partial \mu}{\partial E_p} \div \left(1 + E_g \frac{\partial \mu}{\partial E_p}\right), \quad \mu_o = \frac{\mu - E_p \dfrac{\partial \mu}{\partial E_p}}{1 + E_g \dfrac{\partial \mu}{\partial E_p}}.$$

From the upper curve, for $E_p = 120$, $E_g = -2.5$

$$\mu = 32, \quad \frac{\partial \mu}{\partial E_p} = .05328,$$

whence

$$K = .06146, \mu_o = 29.55$$

We next consider the component of (31) that results in plate curvature detection or modulation. It is given by

$$i_{p2} = (1 - b_{1h}q_h)\ (1 - \bar{b}_{1k}q_k)a_{2(h-k)}\ e_{2(h-k)}$$

$$= \frac{\frac{r_g^2}{2}\left[-\mu^2 r_p r_p' + \mu \frac{\partial \mu}{\partial \bar{E}_p}(r_p^2 - z_h z_k) + \frac{\partial \mu}{\partial E_g}(r_p + z_h)(r_p + \bar{z}_k) \right] e_{2(h-k)}}{\left[r_g + q_h\left(1 - \frac{\mu}{\nu}\frac{z_h}{r_p + z_h}\right)\right]\left[r_g + q_k\left(1 - \frac{\mu}{\nu}\frac{\bar{z}_k}{r_p + \bar{z}_k}\right)\right]} \tag{35}$$
$$(r_p + z_h)(r_p + \bar{z}_k)(r_p + z_{(h-k)})$$

For rough calculations, μ may be regarded as a constant. For very careful work, this assumption should never be made without first drawing the curves of $\mu - E_g$ and $\mu - E_p$ and verifying the validity of the assumption under operating conditions. Examples of such curves are given in Fig. 2. When μ may be regarded as constant, and when the grid is maintained negative with respect to the filament, (35) becomes

$$i_{p(h-k)} = \frac{-\frac{1}{2}\mu^2 r_p r_p' e_{2(h-k)}}{(r_p + z_h)(r_p + \bar{z}_k)(r_p + z_{h-k})}$$

which may be put into the form given in Mr. Carson's paper, referred to before.

The third and last component of (31) is that which produces grid detection or modulation; namely

$$i_{p(h-k)} = -a_{1(h-k)}q_{(h-k)}b_{2(h-k)}e_{2(h-k)}$$

$$= \left(\frac{\mu q_{(h-k)}}{r_p + z_{(h-k)}}\right)\frac{1}{2}\left[-r_g r_g'\left(1 - \frac{\mu}{\nu}\frac{z_h}{r_p + z_h}\right)\left(1 - \frac{\mu}{\nu}\frac{\bar{z}_k}{r_p + \bar{z}_k}\right)\right.$$

$$-\frac{2a_{2(h-k)}r_g^2 z_{(h-k)}}{\nu} + \frac{\partial}{\partial E_p}\left(\frac{1}{\nu}\right)\left(\frac{r_g^2 \mu^2 z_h \bar{z}_k}{(r_p + z_h)(r_p + \bar{z}_k)}\right) - \frac{\partial}{\partial E_g}\left(\frac{1}{\nu}\right)$$

$$\left.\left(\frac{\mu z_h r_g^2}{r_p + z_h} + \frac{\mu \bar{z}_k r_g^2}{r_p + \bar{z}_k} - \frac{\mu^2 r_g^2 z_h \bar{z}_k}{\nu(r_p + z_h)(r_p + \bar{z}_k)}\right)\right]$$

$$\div \left[r_g + q_h\left(1 - \frac{\mu}{\nu}\frac{z_h}{r_p + z_h}\right)\right]\left[r_g + \bar{q}_k\left(1 - \frac{\mu}{\nu}\frac{\bar{z}_k}{r_p + \bar{z}_k}\right)\right]$$

$$\left[r_g + q_{(h-k)}\left(1 - \frac{\mu}{\nu}\frac{z_{(h-k)}}{r_p + z_{h-k}}\right)\right] \tag{36}$$

In using this relation, ν may nearly always be considered constant. As grid leak detectors are often used, q consists of a resistance, R_g, and a condenser, C, in parallel. The values of R_g and C are so ad-

justed that the impedance of the combination to the first order frequencies is practically that of the condenser alone and may be neglected, and to the desired second order, or detected, frequency it is practically that of the resistance alone. When this is the case, and when the impedance in the plate circuit is a pure resistance, R_p, we may write (36) as follows:

$$i_{p3}=\frac{1}{2}\frac{\mu R_g\left(r_g r_g'+\dfrac{2a_{2m}r_g^2R_p}{\nu}\right)}{(r_p+R_p)r_g^2(r_g+R_g)}e_{2m}$$

and, considering μ constant, in order to obtain a physical view of the result, we get

$$i_{p3}=\frac{\dfrac{1}{2}\mu R_g\left[r_g r_g'-\dfrac{\mu^2 r_p r_p'}{(r_p+R_p)^3}\left(\dfrac{r_g^2R_p}{\nu}\right)\right]e_{2m}}{(r_p+R_p)r_g^2(r_g+R_g)} \tag{37}$$

This equation shows a condition that is present in many grid-leak detectors and which, it is thought, has not been generally appreciated. The condition referred to is the presence of the term involving the curvature of the plate characteristic in the grid detection component. This effect is *in addition* to the plate detection effect, given by (35). The plate detection component and the grid detection component are opposite in phase. Hence, it would seem that for best operation as a grid-leak detector, the curvature of the plate characteristic should be zero. This means a rather large value of E_b, the plate battery potential. In practice, however, it is usual to operate with fairly low values of E_b. The second term of the numerator of (37) accounts for this. It will be seen that detection resulting from this term and from the first term are in phase, since ν is intrinsically negative. Hence, it is entirely possible in certain cases for the optimum operating point to be such that the effect of the plate curvature is appreciable.

We now combine once more the three components, (33), (35) and (36), under the simplifying assumptions that μ and ν are constant and that ν is large enough so that terms containing ν in the denominator may be neglected. The result is

$$\left.\begin{aligned} i_p=&\frac{r_g}{(r_g+q_h)}\frac{e_{1h}}{(r_p+z_h)}+\frac{r_q}{(r_q+q_k)}\frac{e_{1k}}{(r_p+z_k)}+\cdots\\ &+\left\{\frac{r_g^2}{(r_g+q_k)(r_g+q_k)}\left[\frac{-\frac{1}{2}\mu^2 r_p r_p'}{(r_p+z_h)(r_p+\bar{z}_k)(r_p+z_{h-k})}\right]\right.\\ &\left.+\frac{\mu q_{(h-k)}}{(r_p+z_{h-k})}\left[\frac{\frac{1}{2}r_g r_g'}{(r_g+q_h)(r_g+\bar{q}_k)(r_g+q_{h-k})}\right]\right\}e_{2(h-k)}+\cdots \end{aligned}\right\} \tag{38}$$

The first two terms of (38) are the amplification terms and represent undistorted reproduction in the plate circuit of the voltage, e, applied in the grid circuit. The third term of (38) represents the second order effects resulting from the curvature of the characteristics of the vacuum tube. The first part of this term represents the effects of so-called plate curvature detection or modulation, and the second part represents the effects of detection and modulation in the grid circuit. It is with this last-named component that the present paper is most concerned.

The Grid-Leak Detector

Fig. 3 shows the usual circuit diagram for a grid-leak detector. It is evident that the impedance, q, in this example is composed of the parallel combination of R_g and C. Suppose that the "h" and "k" frequencies are both radio frequencies, and that, for them, the

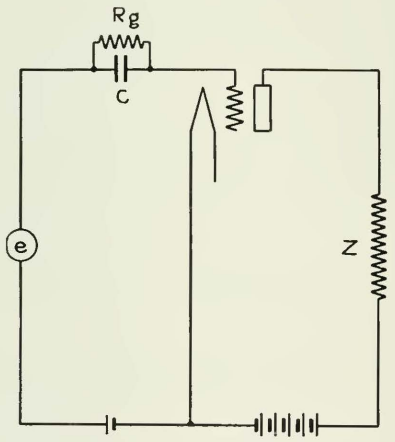


Fig. 3—Grid-leak detector

impedance offered by the resistance and condenser combination is practically that of the condenser, alone. Suppose, further, that practically the only impedance offered by the external circuit to the "$(h\text{-}k)$" frequency is that of the resistance, R_g, alone. This, of course, assumes that the "h-k" frequency is quite low. Then, when E_b, the voltage of the plate battery, is such that r_p' is very small, and when

$$e = A \cos ht + B \cos kt$$

we have, from (38), for second order effects

$$\left.\begin{aligned} i_p = \tfrac{1}{2} r_g r_g' \left(\frac{\mu}{r_p + R_p}\right) \Bigg[& \frac{R_g}{\left(r_g + \frac{1}{jhC}\right)\left(r_g - \frac{1}{jhC}\right)(r_g + R_g)} \frac{A^2}{2} \\ & + \frac{\frac{1}{2jhC}}{\left(r_g + \frac{1}{jhC}\right)^2 \left(r_g + \frac{1}{2jhC}\right)} \frac{A^2}{2} \cos 2ht \\ & + \frac{R_g}{\left(r_g + \frac{1}{jkC}\right)\left(r_g - \frac{1}{jkC}\right)(r_g + R_g)} \frac{B^2}{2} \\ & + \frac{\frac{1}{2jkC}}{\left(r_g + \frac{1}{jkC}\right)^2 \left(r_g + \frac{1}{2jkC}\right)} \frac{B^2}{2} \cos 2kt \\ & + \frac{\frac{1}{j(h+k)C}}{\left(r_g + \frac{1}{jhC}\right)\left(r_g + \frac{1}{jkC}\right)\left(r_g + \frac{1}{j(h+k)C}\right)} AB \cos (h+k)t \\ & + \frac{R_g}{\left(r_g + \frac{1}{jhC}\right)\left(r_g - \frac{1}{jkC}\right)(r_g + R_g)} AB \cos (h-k)t \Bigg] \end{aligned}\right\} \quad (39)$$

While most of the frequencies in this expression are unimportant in relation to any practical case on hand, they are included here to show the complete result for a given simple case. The last term of the above expression results in what is known as detection.

Let us consider this component in more detail as regards detection of an incoming modulated radio wave of the form

$$e = A\ (1 + B \cos qt) \cos pt. \quad (40)$$

They may be written

$$e = A \cos pt + \frac{AB}{2} \cos (p+q)t + \frac{AB}{2} \cos (p-q)t. \quad (41)$$

If we identify the "p" frequency with "h," and let "k" have the values $(p+q)$ and $(p-q)$ in turn, the detection term of (39) gives

$$i_d = \tfrac{1}{2} r_g r_g' \left(\frac{\mu}{r_p + R_p}\right) \left[\frac{R_g}{\left(r_g - \frac{1}{jpC}\right)\left(r_g + \frac{1}{j(p+q)C}\right)(r_g + R_g)} + \frac{R_g}{\left(r_g + \frac{1}{jpC}\right)\left(r_g - \frac{1}{j(p-q)C}\right)(r_g + R_g)} \right] \frac{A^2 B}{2} \cos qt. \quad (42)$$

Reference to the mathematical digression will make clear the formation of the impedances in this expression. (42) is an important relation since it shows that there is a possibility that the amplitude of the detected current may be affected by the phase displacements of the side bands of the original wave which occur during the detection. For an ideal grid-leak detector, the magnitudes of the quantities $\frac{1}{pC}$, $\frac{1}{(p+q)C}$ and $\frac{1}{(p-q)C}$ are very small compared with r_g. Equation (42) then becomes

$$i_d = \frac{r_g'}{r_g} \frac{\mu}{(r_p+R_p)} \frac{R_g}{(r_g+R_g)} \frac{A^2B}{2} \cos qt. \tag{43}$$

In (43) we have the simplest possible form of the equation for a grid leak detector. The next step is to show methods for evaluating the quantities r_g and r_g'. As may be seen from the relations given in (7) and (8)

$$\frac{1}{r_g} = \frac{\partial I_{go}}{\partial E_g}, \qquad r_g' = \frac{\partial r_g}{\partial E_g},$$

and, since the action of the grid-leak detector depends upon r_g', it is evident that r_g is not a constant but varies with the value of E_g. We may obtain r_g by direct dynamical measurements or by drawing tangents to the static grid-potential grid-current curve of the tube under consideration. The value of r_g thus obtained applies only to a given value of E_g. Now E_g is a function of the voltage, e, as will be shown:

When e has the form given in (41), one of the resulting currents in the plate circuit is a direct current given by

$$i_{pd} = \frac{1}{2} \frac{r_g'}{r_g} \frac{\mu R_g}{(r_g+R_g)(r_p+R_p)} \left(\frac{A^2}{2} + \frac{A^2B^2}{2}\right).$$

This means that a constant voltage given by

$$e_{gd} = \frac{1}{2} \frac{r_g'}{r_g} \frac{R_g}{(r_g+R_g)} \left[\frac{A^2}{2} + \frac{A^2B^2}{2}\right] \tag{44}$$

must have appeared on the grid in order to produce the constant component of the plate current. This constant voltage is in addition to that which we have denoted by E_{go}, since it is part of e_g. Moreover, its intrinsic value is usually negative, since r_g' is usually negative. This means that the "effective" E_{go} has been reduced by the amount given in (44). However, r_g is slightly different at this new value of E_{go} and hence e_{gd} is not quite what a first calculation would

lead one to believe. The method of arriving at the correct value for e_{gd}, and hence for r_g and r_g' is one of trial and error, for, after several recalculations of e_{gd} have been made, it will be found that check results are secured. Then r_g and r_g' may be determined from this resulting value of E_{go}.

In actually making these measurements, a dynamical method of measuring r_g will usually be found superior to the method of drawing tangents to the static characteristic, for the grid-potential grid-current characteristic of any tube is rather elusive because of the

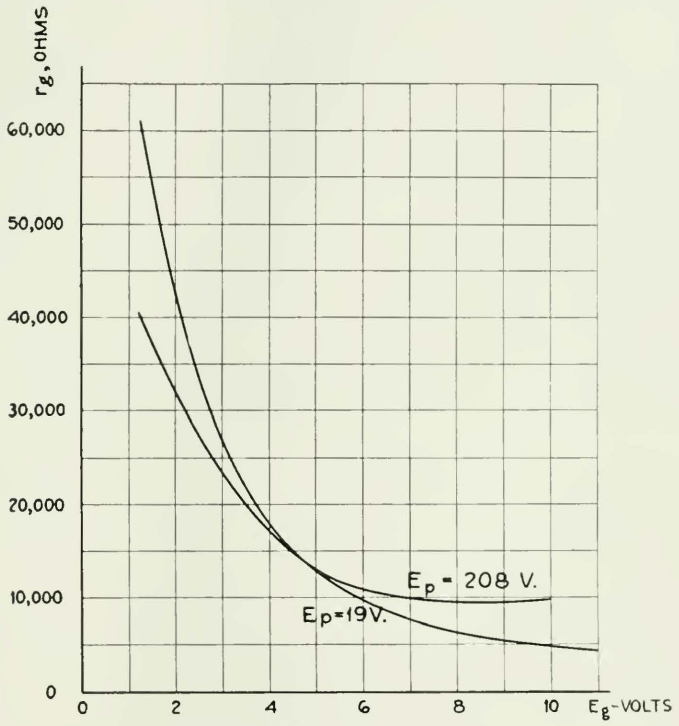


Fig. 4—Grid resistance

very small values of current involved. In the dynamic method a Wheatstone bridge circuit excited by a high frequency buzzer will be found convenient. The value of r_g' is, of course, obtained by drawing tangents to the r_g curve. Several examples of $r_g - E_g$ curves are shown in Fig. 4.

It must be recognized that, for large values of buzzer excitation, the dynamic value of r_g differs somewhat from that found by drawing tangents to the static characteristic. The dynamic value more nearly

approaches the value r_g would assume with large signal inputs than does the static value. Hence, if a large signal input, e, is to be used, the amplitude of the buzzer excitation voltage on the grid should equal this amplitude as nearly as possible.

When the method of drawing tangents to the static characteristic is employed, a very close approximation to the value of r_g to use for large signal amplitudes may be obtained by drawing, not true tangents but secant lines to the static characteristic, which join points on the characteristic corresponding to the extreme, or peak, values of e_g.

When either method is used to obtain r_g, the value of r_g' must be obtained by drawing tangents to an $E_g - r_g$ curve.

With the precautions just given, and when the assumptions made in equation (43) are justifiable, an accuracy within 10% is easily obtained. While this is not very exact, nevertheless, it is a real advance over calculations made without taking the precautions just discussed for measuring r_g.

In many vacuum tubes the value of r_g is so high that the input impedance of the tube, resulting from the interelectrode capacities of the elements cannot justifiably be neglected. In order to include this effect, the following relations are applicable.

Consider the circuits shown in Fig. 5. This gives the equivalent circuit diagram for a vacuum tube with general impedances, z_1 and

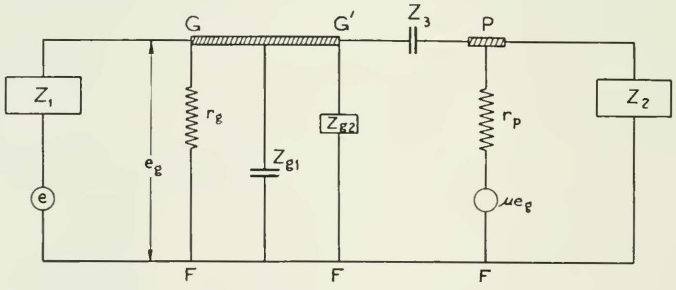


Fig. 5—Equivalent network

z_2, attached to the grid and plate, respectively. The plate to filament capacity may conveniently be included in z_2. The impedance, z_{g_2}, is the effective impedance of the network looking to the right from the point $G'F$. z_{g_1} is the grid to filament capacity of the tube, and z_3 is the grid to plate capacity.

We may write

$$z_g = \frac{z_{g_1} z_{g_2}}{z_{g_1} + z_{g_2}} .$$

In order to apply the general equations we must evaluate z_n and q_n.

To Evaluate z_n.

From the general equations, we have

$$i_p = \frac{\mu e_g}{r_p + z_n}.$$

Hence we may write Kirchoff's law for the plate circuit. This gives

$$i_p = \frac{e_g[(\mu+1)z_2 + \mu z_3]}{r_p z_2 + z_3(r_p + z_2)}.$$

Upon equating the two expressions for i_p, there results

$$z_n = \frac{z_2(\mu z_3 - r_p)}{(\mu+1)z_2 + \mu z_3}.$$

To Evaluate q_n.

By the general equations, we have

$$i_g = \frac{e}{r_g + q_n x}$$

where x stands for

$$\left(1 - \frac{\mu}{\nu}\frac{z_n}{r_p + z_n}\right).$$

This may be written

$$i_g = \frac{\frac{e}{x}}{\frac{r_g}{x} + q_n}$$

which says that Kirchoff's law may be applied to the grid circuit provided we use a modified voltage, $\frac{e}{x}$, and a modified grid resistance, $\frac{r_g}{x}$. Hence

$$i_g = \frac{\frac{e}{x} z_g}{(z_1 + z_g)\left(z_g + \frac{r_g}{x}\right) - z_g^2}.$$

Upon equating the two expressions for i_g there results

$$q_n = \frac{z_1}{z_g}\left(\frac{r_g}{1 - \frac{\mu}{\nu}\frac{z_n}{r_p + z_n}} + z_g\right).$$

To sum up; the following relations are applicable when interelectrode capacities or other coupling impedances are to be included:

$$z_g = \frac{z_{g1} z_{g2}}{z_{g1} + z_{g2}} \tag{45}$$

$$z_{g1} = \frac{1}{j\omega C_{gf}} \tag{46}$$

$$z_{g2} = \frac{z_2 z_3 + r_p(z_2 + z_3)}{z_2(\mu + 1) + r_p} \tag{47}$$

$$z_n = \frac{z_2(\mu z_3 - r_p)}{(\mu + 1)z_2 + \mu z_3} \tag{48}$$

$$q_n = \frac{z_1}{z_g}\left(\frac{r_g}{1 - \frac{\mu}{\nu}\frac{z_n}{r_p + z_n}} + z_g \right). \tag{49}$$

With the aid of (45), (46), (47), (48), (49), equation (42) may be modified to include all cases where the plate current resulting from detection or modulation in the grid circuit is desired, provided an accuracy greater than about 10% is not required. Where greater accuracy is essential, curves must be made to give the effect of the small terms in the numerator of the expression for b_{2m} in equation (36).

Before leaving the subject of grid-leak detectors, we will discuss briefly one of the physical aspects of grid-leak detection that the example just given, and the equations on which it is based, have emphasized. This is the fact that the fiction of the time-constant of the grid-leak and condenser combination is not a necessary physical interpretation of the phenomena which occur in the grid circuit. Indeed, in many cases, the time constant method of calculating the leak and condenser gives quite erroneous and misleading results. These cases occur when the impedance looking into the vacuum tube is of such value, as it often is, that the magnitudes and forms of q_{1n} and q_{2m} are materially changed from those which they would have if z_g were neglected, and when r_g is not large compared with q_n and q_m. Equation (38) shows that, for greatest plate current resulting from grid detection, q_h and q_k should be as small as possible, while $q_{(h-k)}$ should be as large as possible. It is, then, a filter problem, and if treated as such, will give reliable results both as to physical interpretation and numerical values.

In the special case when the input and detected frequencies are $\frac{h}{2\pi}$ and $\frac{s}{2\pi}$, respectively, and where $z_g \gg r_g$:

$$q_h = \frac{1}{jhC}$$

$$q_s = \frac{\frac{R}{jsC}}{R + \frac{1}{jsC}}$$

R = leak resistance

C = capacity in parallel with R

Then, the optimum size for the condenser, C, is easily shown to be

$$C^2 = \frac{\sqrt{2}\,(R+r_g)}{hsRr_g^2}, \text{ (approx.)} \tag{50}$$



Fig. 6—Optimum size of grid-leak condenser

Experimental Conditions:

$h = 2\pi \times (30000 \pm 500)$

$s = 2\pi \times (1000)$

Grid-leak $= R = 10^6$ ohms

$r_g = 10^5$ ohms

Calculation Conditions:

$z_g \gg r_g$

$$q_h = \frac{i}{jhC}$$

$$q_s = \frac{\frac{R}{jsC}}{R + \frac{i}{jsC}}$$

Then the optimum size of the grid-leak condenser, C, is:

$$C_{opt}^2 = \frac{\sqrt{2}\,(R+r_g)}{hsR\,r_g^2}$$

or:

$$C_{opt} = 361\mu\mu \text{ farad}$$

Fig. 6 illustrates the agreement between this relation and an actual circuit where the above conditions were closely approximated.

Plate Curvature Detection

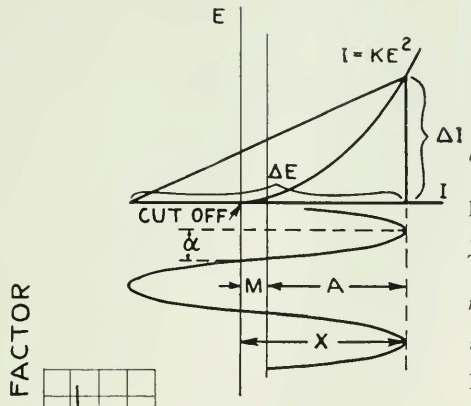
In discussing this phase of the problem we refer to equation (35). In addition to the remarks made in connection with that equation it is necessary only to add a few words on the evaluation of r_p and r_p'. In general, these quantities are susceptible to the same method of treatment that was suggested in dealing with r_g and r_g'. Two fundamental circuits for plate curvature detectors are in use. In the first the plate battery is placed in series with the load impedance. In the case when the load impedance contains appreciable resistance the normal or effective value of E_p must be obtained in the manner described for finding E_g. In the second circuit the plate battery potential is introduced through a low resistance, high impedance, choke, and the normal value of E_p is then equal to E_b. Especially in dealing with resistance coupled units these points should be borne in mind.

Amplification

Equation (33) gives the general amplification relation. The remarks made under the heading of the "Grid-Leak Detector" concerning the evaluation of the z's and q's are applicable here, as in all other vacuum tube relations. The special points to be brought out are the methods of applying the equations to so-called improper amplifiers of Class III. In this type of amplifier the grid swings negative further than the plate current cut-off point each cycle. Experience has shown that even in this event, to find the tube resistances, the approximation of using the secant line joining two points on the characteristic corresponding to the extreme values of the input voltage, is often justifiable. If greater accuracy is desired, the corrections given by the curve, Fig. 7, should be applied. These corrections are based on the assumption of a sine wave input and a characteristic that follows the square law, and to that extent are themselves in error. For modulated waves the dotted curves give values found by interpolation between the two points shown.

Modulation

The detection equations apply equally well to modulation effects. The only case in which a question may arise is that in which one of the input frequencies is introduced into the plate circuit of the tube



Explanation:

$I = KE^2$

$= K(M + A \cos pt)^2|_{E+}$

$= \frac{b_0}{2} + b_1 \cos pt + b_2 \cos 2\ pt + \ldots$

$b_1 = \frac{2KA^2}{\pi}\left[\frac{3}{4}\sin\alpha - \alpha\cos\alpha + \frac{1}{12}\sin 3\alpha\right]$

Fundamental current is:

$i_f = b_1 \cos pt.$

Then

$r = \frac{e}{i} = \frac{A \cos pt}{b_1 \cos pt} = \frac{A}{b_1}$

and

$\text{Factor} = \frac{r}{\Delta E / \Delta I}.$]

Hence the resistance to pure sine wave of square-law devices in the vicinity of cut-off is

$$r = \frac{\Delta E}{\Delta I} \times \text{Factor.}$$

Dotted curves show factor for modulated waves of the form $A(1+B \cos qt) \cos pt$ where

$$\Delta E = 2A(1+B).$$

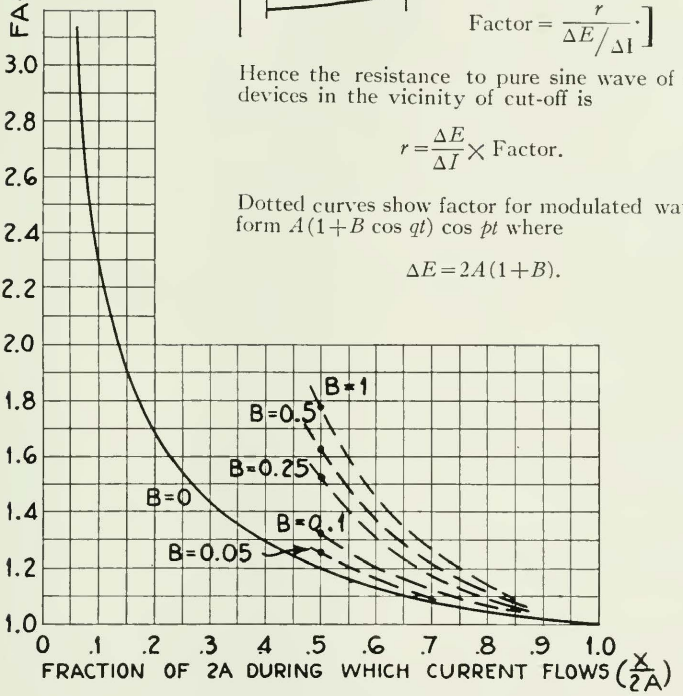


Fig. 7—Correction factor for resistance of non-linear device

while the other is introduced into the grid circuit. To analyze this condition for the general case, (see Fig. 8) let lower case e's refer to the driving voltage impressed directly on the grid. Let the E's refer to the driving voltage in series with an impedance in the plate circuit. We then have the series

$$i_p = a_1(E+e) + a_2(E+e)^2 + \ldots$$

which, in accordance with the complex quantity notation may be written

$$i_p = a_{1h}E + a_{1k}e + a_{2(2h)}E^2 + a_{2(2k)}e^2 + 2a_{2(OE)}E\bar{E} + 2a_{2(h+k)}Ee + 2a_{2(h-k)}E\bar{e} + 2a_{2(Oe)}e\bar{e} + \ldots$$

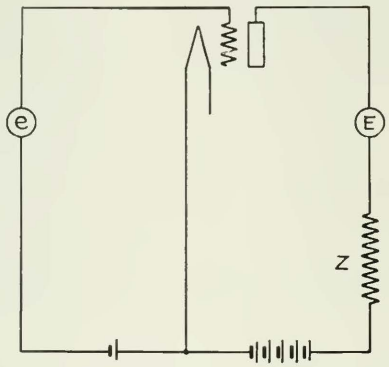


Fig. 8—Plate circuit modulation

Then, with the aid of (4), upon equating coefficients of like powers of e, E, and Ee, we get

$$a_{1h} = \frac{1}{r_p + z_h} \qquad a_{1k} = \frac{\mu}{r_p + z_k}$$

$$a_{2(2h)} = \frac{-\frac{1}{2}r_p r_p'}{(r_p + z_h)^2(r_p + z_{2h})}$$

$$a_{2(2k)} = \frac{\frac{1}{2}\left[\frac{\partial \mu}{\partial E_g}(r_p + z_k)^2 + \mu\frac{\partial \mu}{\partial E_p}(r_p^2 - z_k^2) - \mu^2 r_p r_p'\right]}{(r_p + z_k)^2(r_p + z_{2k})}$$

$$a_{2(OE)} = \frac{-\frac{1}{2}r_p r_p'}{(r_p + z_h)^2(r_p + R)} \qquad a_{2(h+k)} = \frac{-\frac{1}{2}\left[\frac{\partial \mu}{\partial E_p}\frac{r_p}{2}(2r_p + z_h + z_k) - \mu r_p r_p'\right]}{(r_p + z_h)(r_p + z_k)(r_p + z_{h+k})}$$

$$a_{2(h-k)} = \frac{\frac{1}{2}\left[\frac{\partial \mu}{\partial E_p}\frac{r_p}{2}(2r_p + z_h + \bar{z}_k) - \mu r_p r_p'\right]}{(r_p + z_h)(r_p + z_k)(r_p + z_{h-k})} \tag{51}$$

$$a_{2(Oe)} = \frac{\frac{1}{2}\left[\frac{\partial \mu}{\partial E_g}(r_p + z_k)^2 + \mu\frac{\partial \mu}{\partial E_p}(r_p^2 - z_k)^2 - \mu^2 r_p r_p'\right]}{(r_p + z_k)^2(r_p + R)}$$

When z is a resistance, R, the expression for i_p reduces to

$$i_p = \frac{(\mu e+E)}{r_p+R} - \frac{\frac{1}{2}r_p r_p'}{(r_p+R)^3}E^2 + \frac{\frac{1}{2}\left[\frac{\partial \mu}{\partial E_g}(r_p+R)^2+\mu\frac{\partial \mu}{\partial E_p}(r_p^2-R^2)-\mu^2 r_p r_p'\right]}{(r_p+R)^3}e^2 + \frac{\frac{\partial \mu}{\partial E_p}r_p(r_p+R)-\mu r_p r_p'}{(r_p+R)^3}Ee+ \text{- - - -} \tag{52}$$

If μ is constant, this becomes

$$i_p = \frac{\mu e+E}{r_p+R} - \frac{\frac{1}{2}r_p r_p'}{(r_p+R)^3}(\mu e+E)^2+ \text{- - -} \tag{53}$$

which shows that the circuit then acts as though a voltage, $(\mu e+E)$ had been impressed in series with the plate circuit.

Oscillation

The subject of vacuum tube oscillators has been so extensively treated elsewhere that but little new material has thus far been obtained from the general equations now offered. The method of handling the problem is, however, illuminating as it gives an example of what is meant by the statement that no sharply drawn line should be placed between oscillation, detection, amplification, or other uses of the thermionic vacuum tube.

In treating the oscillator problem we consider the amplification term of the general equations; namely

$$i_p = \frac{\mu e}{(r_p+z_n)}\frac{r_g}{\left[r_g+q_n\left(1-\frac{\mu}{\nu}\frac{z_n}{r_p+z_n}\right)\right]}.$$

The oscillating conditions require that current shall flow without a driving voltage. Hence, as e is zero, i_p can be finite only if one of the factors in the denominator is zero. Thus either

$$r_p+z_n=0 \tag{54}$$

or

$$r_g+q_n\left(1-\frac{\mu}{\nu}\frac{z_n}{r_p+z_n}\right)=0 \tag{55}$$

gives the conditions for oscillation. Fig. 5 and the relations of (45), (46), (47), (48) and (49) are applicable here. The condition of (54) requires a negative value of r_p, and hence is not the usual oscillation condition. The condition of (55) therefore gives the criterion for the oscillation condition. As before, neglecting quantities in $\frac{1}{\nu}$ we may write (55) in the following form

$$r_g + q_n = 0$$

or

$$r_g + \frac{z_1(z_g + r_g)}{z_g} = 0. \tag{56}$$

When applied to a hypothetical Hartley oscillator, Fig. 9, with the circuit constants

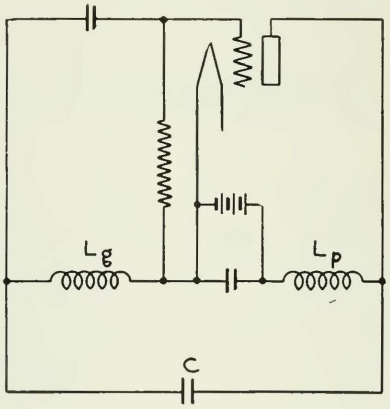


Fig. 9—Hartley oscillator

$$z_2 = j\omega L_p \qquad z_3 = \frac{1}{j\omega C} \qquad z_1 = j\omega L_g,$$

equation (56) gives as the conditions for oscillation

$$\omega^2 = \frac{1}{\left[L_p + L_g + \frac{L_p L_g}{C r_p r_g}\right]} = \frac{1}{(L_p + L_g)C} \text{ (nearly)}, \tag{57}$$

$$L_p = \left[\mu - \frac{r_p}{r_g L_p}\right] L_g = \mu L_g \text{ (nearly)}. \tag{58}$$

The relations of (57) and (58) have been given many times, and are included here only in order to illustrate the ease with which simple problems may be solved from fundamental relations.

Application of the Theory

The illustrations will serve to give a sufficiently comprehensive view of the methods of applying the general equations to special cases.

Inasmuch as the derivation of the equations requires no assumptions other than that the static curves of grid current-grid potential, and plate current-plate potential of the tube remain constant, the accuracy with which a given problem may be calculated depends only upon the ability to determine the effective differential coefficients required by the Taylor's series expansions, and the number of terms of the series included. Practically, the component of current of a given frequency resulting from any higher order term is entirely negligible with respect to the component of the same frequency resulting from lower order terms. For precise results in a general case the calculations are necessarily tedious, since the physical processes are quite complex. However, in any given special case one of the respective approximations indicated is usually allowable, which greatly simplifies matters. In the event that any question arises concerning the proper phase angles for the complex impedances, the correct result may always be arrived at by writing the voltages in full complex form, as illustrated in the mathematical digression. The impedances will then take care of themselves.

While it is difficult to show mathematically the convergence of the series of (31), experience has shown that the convergence is so rapid that higher order terms may be neglected, unless new frequencies developed by them are under investigation. In these cases, the conditions of the problem are often such that simplifying assumptions may be made at the outset. If familiarity with the complex impedances has been attained, it will, in many cases, be sufficient to derive all equations on the basis of resistance only, and then introduce the complex impedances in the manner indicated by the analogy between these and the general equations.

The higher order coefficients are given below for the special case where resistances, only, are considered, and where the voltage, e_g, is known. It is found more convenient to use the P's, equation (4), in their derivative form than to attempt to express them in terms of μ and r_p, so referring to the expansion

$$i_p = a_1 e_g + a_2 e_g^2 + a_3 e_g^3 + a_4 e_g^4 + a_5 e_g^5 + _ \cdot \cdot \cdot,$$

we have

$$a_1 = \frac{P_1}{1+P_2Z}$$

$$a_2 = \frac{\frac{1}{|2}}{1+P_2Z}\left[P_3 - 2P_4\frac{P_1Z}{1+P_2Z} + P_5\frac{Z^2P_1^2}{(1+P_2Z)^2}\right],$$

$$a_3 = \frac{1}{1+P_2Z}\left(\frac{1}{|2}\left[-2P_4a_2Z + 2P_5a_1a_2Z_1Z_2\right]\right.$$

$$\left. + \frac{1}{|3}\left[P_6 - 3P_7a_1Z + 3P_8a_1^2Z^2 - P_9a_1^3Z^3\right]\right),$$

$$a_4 = \frac{1}{1+P_2Z}\left(\frac{1}{|2}\left[-2P_4a_3Z + P_5(a_2^2Z^2 + 2a_1a_3Z_1Z_3\right]\right.$$

$$+ \frac{1}{|3}\left[-3P_7a_2Z + 6P_8a_1a_2Z^2 - 3P_9a_1^2a_2Z^3\right]$$

$$\left. + \frac{1}{|4}\left[P_{10} - 4P_{11}a_1Z + 6P_{12}a_1^2Z^2 - 4P_{13}a_1^3Z^3 + P_{14}a_1^4Z^4\right]\right)$$

$$a_5 = \frac{1}{1+P_2Z}\left(\frac{1}{|2}\left[-2P_4a_4Z + 2P_5(a_1a_4Z^2 + a_2a_3Z^2)\right]\right.$$

$$+ \frac{1}{|3}\left[-3P_7a_3Z + 3P_8(a_2^2Z^2 + 2a_1a_3Z^2)\right.$$

$$\left. - P_9(a_1Za_2^2Z^2 + 2a_2a_3Z^2 + 2a_1a_2^2Z^3 + a_1^2a_3Z^3)\right]$$

$$+ \frac{1}{|4}\left[-4P_{11}a_2Z + 12P_{12}a_1a_2Z^2 - 12P_{13}a_1^2a_2Z^3 + 4P_{14}a_1^3a_2Z^4\right]$$

$$+ \frac{1}{|5}\left[P_{15} - 5P_{16}a_1Z + 10P_{17}a_1^2Z^2 - 10P_{18}a_1^3Z^3\right.$$

$$\left. \left. + 5P_{19}a_1^4Z^4 - P_{20}a_1^5Z^5\right)\right],$$

APPENDIX I

To Show that with Negative Grid Potentials the Relation:

$$\mu \frac{\partial \mu}{\partial E_p} = \frac{\partial \mu}{\partial E_g}$$

Holds With Fair Precision

We have the fundamental expression:

$$I_p = I_p(E_g, E_g) \tag{1}$$

Suppose that E_g and E_p are allowed to vary under the restriction that I_p is maintained constant. Then:

$$d\, I_p = 0 \tag{2}$$

Hence:

$$\frac{dI_p}{dE_g} = 0 = \frac{\partial I_p}{\partial E_g} + \frac{\partial I_p}{\partial E_p} \frac{dE_p}{dE_g} \tag{3}$$

Whence:

$$-\frac{dE_p}{dE_g}\Big]_{I_p} = -\mu \tag{4}$$

Also:

$$d^2 I_p = 0 \tag{5}$$

Hence:

$$\frac{d^2 I_p}{dE_g^2} = 0 = \frac{\partial^2 I_p}{\partial E_g^2} + 2\frac{\partial^2 I_p}{\partial E_g \partial E_p}\frac{dE_p}{dE_g} + \frac{\partial^2 I_p}{\partial E_g^2}\left(\frac{dE_p}{dE_g}\right)^2 + \frac{\partial I_p}{\partial E_p}\frac{d^2 E_p}{dE_g^2} \tag{6}$$

Then with the aid of (4), above, and (6) in the body of the paper, we get

$$\frac{\partial \mu}{\partial E_g} - \mu \frac{\partial \mu}{\partial E_p} + \frac{d^2 E_p}{dE_g^2} = 0 \tag{7}$$

Equation (7) shows that:

$$\frac{\partial \mu}{\partial E_g} = \mu \frac{\partial \mu}{\partial E_p} \tag{8}$$

provided that:

$$\frac{d^2 E_p}{dE_g^2} = 0 \tag{9}$$

when I_p is constant.

Experimental curves showing the relation between E_p and E_g required to maintain I_p constant are straight lines, to a very close

approximation, in the region where the grid potential is negative with respect to the filament as shown in Fig. 10. Hence, in this region (9) is satisfied for all practical purposes, and, therefore, the proof of (8) follows directly.

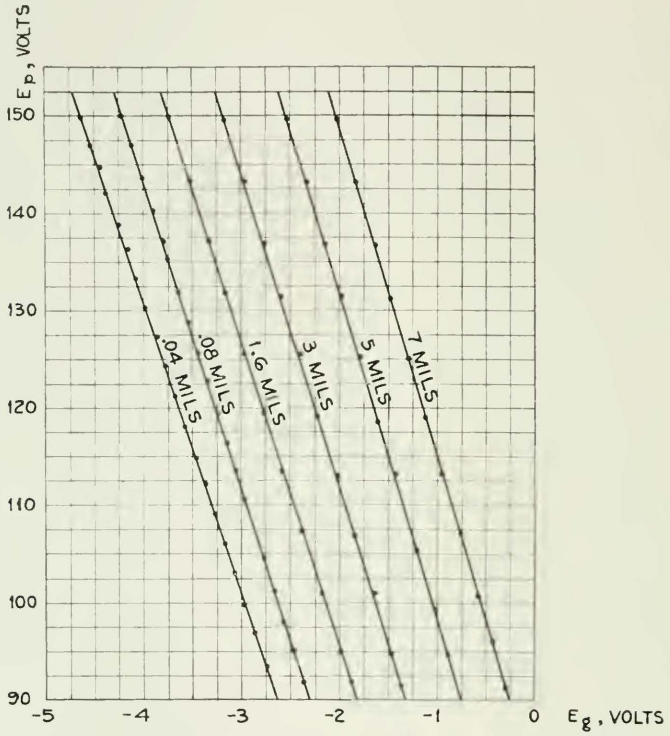


Fig. 10—Relation between E and E for constant plate current

Contemporary Advances in Physics—XI
Ionization

By KARL K. DARROW

IONIZATION, in its most general sense, signifies a segregation of positive from negative charge within the volume of a substance which as a whole is (or initially was) electrically neutral. In practice a gas (for instance) is said to be *ionized* if charges of either sign can be extracted from it. Charged particles of both signs, electrons and ions, can be drawn out from a gas in which an electrical discharge is being maintained; in such a condition, therefore, a gas is ionized. Millikan's droplets, floating around in a gas which had recently been irradiated, absorbed charges of either sign out of the gas, which therefore was ionized by the radiation and remained ionized for some time afterward. A negatively-charged electrode immersed in a carefully-screened gas receives very little charge from it; this condition continues if the gas is bombarded with electrons having less than a certain speed; let the speed of the bombarding electrons be increased past this limit, and the electrode begins to receive positive charge—the gas is ionized by the electrons. Dilute electrolytic solutions are evidently in a continual and spontaneous state of ionization.

Observations on positive ions issuing from ionized gases have been interpreted as meaning that all such ions are atoms or molecules bearing charges of which the magnitude is e, or $2e$, or some other small-integer multiple of e; in other words, as meaning that positive ions are atoms or molecules from which one or more electrons have been detached. Generalizing from these to all cases, it is believed that the first stage of ionization, in monatomic gases at least, is the detachment of electrons from atoms. Whether the separated electrons remain free, or attach themselves to other atoms, or become the gathering-agents of clusters of atoms, is an interesting but at present subsidiary question. Ionization in monatomic gases begins by the detachment of electrons from atoms; and the word "ionization" in fact is frequently used to mean this process alone. In diatomic and compounded gases, the nature of the ions observed permits either of two suppositions; the initial process of ionization may be the detachment of electrons from molecules, or the splitting of molecules into fragments each consisting of one or more atoms, some of these fragments having an excess and the others a compensating deficit of electrons. Special experiments must be performed to decide between these suppositions.

While self-sustaining discharges in gases may produce a vast variety of identifiable ions, they are not suitable for revealing the process of producing these ions. By bombarding a gas with electrons of known speed, ions may be produced under very simple and intelligible conditions. It is then found that in order to detach an electron from an atom of a monatomic gas, a definite amount of energy, the *ionizing-energy* of the gas, must be transferred to the atom. The ionizing-energy is not unique; for most kinds of atoms there are several distinct quantities answering to the same definition. Nevertheless there is one particular and outstanding value which is particularly known as *the* ionizing-energy or *ionizing-potential.* It varies periodically from element to element along the Periodic Table, and is therefore ascribed to an outer electron of the atom; indeed it may be described as the *extraction-energy for the outermost or loosest electron.*

Of the other values of ionizing-energy for a given atom, some are lower than the principal ionizing-potential. These, however, are attributed to atoms in abnormal states. The others are greater than the principal ionizing-potential; some of them are very much greater and increase steadily from one element to the next along the Periodic Table, and are therefore ascribed to deeper-lying electrons and may be described as *extraction-energies for inner electrons.*

The spontaneous ionization of radioactive substances is an entirely irregular function of atomic number and is attributed, for this and other reasons, to events occurring in the nuclei.

At this point it is necessary to define some units. In most determinations of ionizing-energies, a stream of electrons originally moving with speeds thought negligibly small is accelerated by a potential-rise and then projected into the gas under examination. Their kinetic energies in ergs are thus given in terms of the *voltage* V of the potential-rise by the equation

$$\text{Kinetic Energy} = eV/300 = 1.591 \cdot 10^{-12}\ V. \tag{1}$$

It is customary to measure the kinetic energy of an electron by the voltage-rise which gave it, or could have given it, that energy; which is tantamount to employing a unit of energy equal to $1.591.10^{-12}$ erg. This unit may be called the *equivalent volt.*

$$\textit{One equivalent volt} = 1.591 \cdot 10^{-12}\ \textit{erg} \tag{2}$$

The name, it must be admitted, is neither short nor elegant; at all events it is preferable to the slovenly usage of speaking of an electron as having so many "volts of energy" (!) or a "speed of so many volts" (!!) On the other hand, it seems quite unobjectionable to speak of an

electron having a kinetic energy of one equivalent volt as a "one-volt electron."

The ionizing-energy of an atom is usually given in equivalent volts, whence the name *ionizing-potential.*

Occasionally one meets with a value stated for an ionizing-potential in terms of a unit known as the *wave-number* ("equivalent wave-number" would be better) which amounts to $1.968.10^{-16}$ erg.

Ionization-Potentials

The ionizing-potential of a monatomic gas is usually measured by projecting electrons with controllable kinetic energy K into the gas, and determining the value of K at which current begins to flow into an electrode inserted into the gas and maintained at such a potential that positive ions, but no electrons, can reach it.

This method requires more elaborate apparatus than the outline suggests. The experimenter must guard against an effect which was not suspected by those who first worked with the method. Electrons having kinetic energy less the ionizing-energy of the gas may cause the atoms which they strike to emit radiation. Some of this radiation falls upon the electrode arranged to collect positive ions, and expels electrons from it. The field around the collecting-electrode, being such as to draw positive ions toward it, drives these electrons away; and so there is a continuous current of negative charge out of the electrode into the gas, which is quite indistinguishable from a current of positive charge out of the gas into the electrode. Thus the value of K at which positive charge first seems to flow into the electrode from the gas is the "critical" electron-energy (as the phrase is) not for producing ionization but for producing radiation. The earliest determinations of what were thought to be ionizing-potentials were vitiated by this effect.

To avoid or recognize the influence of radiation several schemes have been devised.[1] For example, if two collecting-electrodes are used in alternation, one having a large area and the other being small, much more radiation will fall upon the larger one, and there will be a correspondingly great difference between the currents of negative charge out of the two; but if ions are being formed in the gas, the difference between the numbers of these which find their way to the large and to the small electrode will be much less pronounced. A slender collecting-electrode may record only a very small current due to radiation, but a

[1] For a detailed account of the methods developed up to 1924, consult K. T. Compton and F. L. Mohler: "Critical potentials" (*Bull. Nat. Res. Council*, No. 48).

very large one whenever ionization commences. This scheme has been adopted by K. T. Compton.

Another, and the most common, device for distinguishing ionization from radiation consists in surrounding the collector with a sheath of metal gauze, maintained at a potential slightly (say 3 volts) more negative than the electrode which it screens. Positive ions pass through its meshes to the collector, somewhat slowed down but not driven back. Radiation also passes through the meshes to the collector, but the electrons which it drives out are turned back by the adverse field and reenter the metal whence they came, so that the net result is the same as though they had never come out. Ionization thus produces a current of positive charge into the collector, and radiation none; or radiation may even produce a current of negative charge into the collector, thus accentuating the contrast, for electrons which are ejected from the metal gauze are drawn to the screened electrode. This is the scheme devised by F. S. Goucher.

Another, and possibly the best, method for measuring ionizing-potentials is quite insensitive to radiation. A hot filament is immersed in the gas, which may be supposed to be surrounded by connected metal walls so that its boundaries are all at the same potential. If the efflux of electrons from the filament is so plentiful that it is limited by space-charge,[2] and this condition persists as the potential-difference between walls and filament is raised to the value just sufficing to give to the electrons energy enough to ionize the gas, then at the moment of incipient ionization the space-charge limitation is partially or totally cancelled, and the current increases sharply. This is I. Langmuir's method. It is better to keep the potential-difference between the walls and the filament small and constant, and admit into the gas electrons with controllable energy from another source; when the energy of these auxiliary electrons is raised to attain the ionizing-potential of the gas, the current from the filament suddenly increases. This is the method of G. Hertz[3] and K. H. Kingdon.[4]

Most of the accurate measurements of ionizing-potentials have been made with a collecting-electrode sheathed by a gauze, according to the precept of Goucher. The apparatus is a complicated affair, for the parts already mentioned are by no means all that are required; in some cases the whole interior of the tube appears to be webbed with gauzes. A hot filament (occasionally an illuminated metal plate) is provided as source for electrons, and its potential—or the potential of

[2] See the sixth article of this series (December, 1924).

[3] *ZS. f. Phys.* 18, pp. 307–316 (1923).

[4] *Phys. Rev.* (2) 21, pp. 404–418 (1923).

its negative end—is taken as the zero from which the other potentials are measured. In the sketch (Fig. 1) this is marked F. Close to the source there is a gauze (G_1) maintained at the controllable potential V and thus providing the potential-rise by which the electrons are accelerated. It is clearly desirable that the electrons should move at their

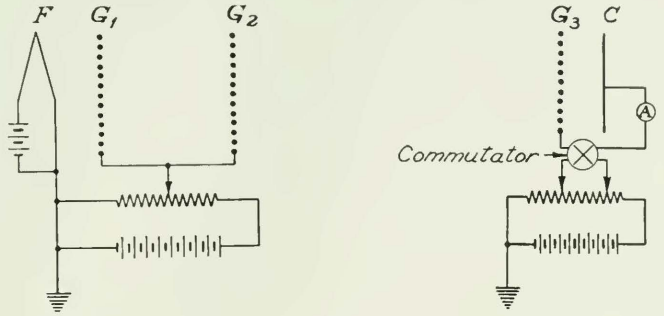


Fig. 1

known maximum speed over as long a path as possible in the gas; consequently a second gauze (G_2) is set up beyond G_1, and maintained at nearly the potential V so that there is a nearly equipotential region between them. (Generally the potential of G_2 is raised a fraction of a volt above V so that there may be a slight impulsion of the ions formed between G_1 and G_2 toward the collector.) Beyond G_2 are the collector C and its protecting gauze G_3, maintained at potentials lower than the filament so that no electrons may reach them. The current of which the sign indicates whether it is due to radiation or ionization, as was explained above, flows through the galvanometer at A.

With such an apparatus as this it seems to be easy enough to measure ionizing-potentials correctly within one or two volts. As soon as greater accuracy is sought after, the real troubles begin. The electrons do not all leave the source with negligible speed; their speeds are distributed over a finite range. The potential to which they climb in passing through the meshes of a gauze is not quite equal to the potential of the gauze-wires themselves. The potential-differences between the different electrodes are not accurately given by voltmeters, for there are contact-potential-differences superposed upon the values indicated. The filament is not an equipotential surface if it is heated by a current, although this difficulty can be overcome if the experimenter thinks it worth the trouble. Electric charges marooned upon the walls of the tube, electrons ejected by radiation from the gauze G_3 and accelerated backwards to G_2 with a final speed higher than the electrons from F

ever attain, are capable of causing false conclusions. The third significant figure in the value of an ionizing-potential is many times harder to attain than the first two; and it is not surprising that many experimenters have chosen to mix some standard gas such as helium into the gases with which they experimented, and to determine the difference between the ionizing-potentials of the standard gas and the other gases, rather than any of them absolutely.

Before bringing out the numerical values of ionizing-potentials, I must allude to the fact that the quantity measured in these experiments is the kinetic energy possessed by the electrons when they are just able to ionize the atoms, which might not be the same thing as the energy actually transferred to the atoms. A particle of mass m moving with speed u has not only kinetic energy $K = \frac{1}{2}mu^2$ but also momentum mu. If it impinges against a previously-stationary particle of mass M, and momentum is conserved in the impact, then the particles must be in motion after the impact, and some of the initial kinetic energy of the striking particle must be saved, so to speak, to provide for this motion. What is left over is available for ionization or other purposes. Without involving ourselves in the general case, we may note that the most favourable conceivable case for having a large proportion of energy left over, when the striking particle is less massive than the struck one, is that in which the more massive particle has all the momentum after the impact. Suppose therefore that after the impact the striking electron and the liberated electron are both stationary, and the ion of mass M is moving with speed V. Conservation of momentum is expressed by writing:

$$mu = MV. \tag{3}$$

The energy T available for ionization or other purposes is given by:

$$K = \tfrac{1}{2}mu^2 = \tfrac{1}{2}MV^2 + T. \tag{4}$$

so that

$$T = K(1 - m/M). \tag{5}$$

Since the masses of atoms range from 1845 to nearly half a million times the mass of an electron, an electron might spend over 999 promille of its energy in ionizing an atom; and therefore there is no essential impossibility in supposing that the energy possessed by an electron just able to ionize is actually equal, within the uncertainty of measurement, to the ionizing-energy of the atom. This supposition is confirmed by the agreements between observed ionizing-potentials and the theoretical values deduced from spectra by using Bohr's method of interpretation.

If the gas or the electron-stream is extremely dense, positive ions appear when the energy of the bombarding electrons is lower than the ionizing-energy as determined by experiments with more rarefied gas or a scantier stream of electrons. Various reasons are assigned for this in various cases; one fundamental reason is, that an atom struck by an electron having less than the ionizing-energy may be put into abnormal states of some duration, in which it can be ionized by receiving a smaller amount of energy than would ionize it in its normal state.

In Fig. 2 the measured values of ionizing-potential are plotted.[5]

There is a way of expressing these and other yet-to-be-presented facts about ionizing-energies, which at this point will probably seem

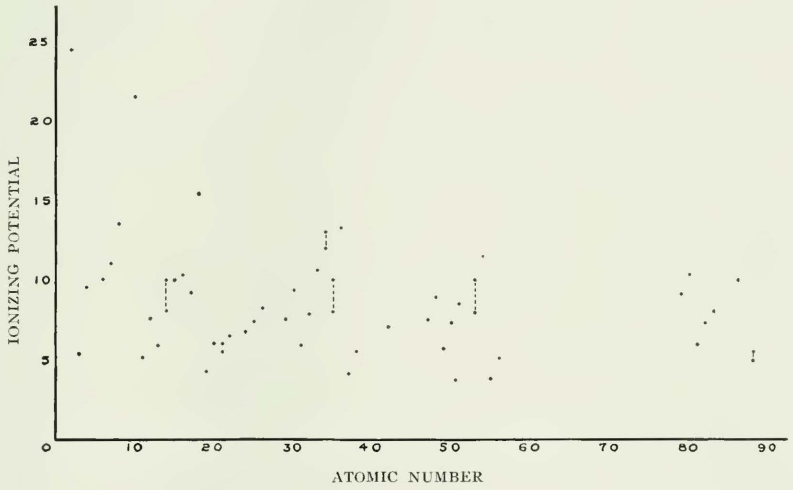


Fig. 2

unnatural but later will be highly convenient. Suppose that by transfer of the ionizing-energy V_0 to an atom it is converted into a system composed of an ion bearing charge $+ e$ and a free electron. This system has potential energy V_0 relatively to the normal state of the atom. The detached electron may wander off and the ion eventually unite itself with another electron. It is convenient, therefore (whether or not it is strictly legitimate) to think of this potential energy V_0 as being associated with the ion alone; and to say that the atom possesses, in addition to its normal state, one or more *states of ionization* or *states of the ionized atom*, each of them characterized by a certain value of po-

[5] I am deeply indebted to Professor F. A. Saunders, who has kept a current catalogue of published values of ionizing-potentials, for enabling me to copy his tabulations.

tential energy. The ionizing-potential of the atom is then, by definition, equal to the potential energy of that state of ionization which differs least in energy from the normal state.

Another way of describing the ionizing-potential is to say that it is the *energy required to detach the loosest electron from the atom.* This involves a picture of an atom as a system of separately-identifiable electrons, "bound" with various degrees of looseness or tightness. Such a picture is so nearly indispensable, that there need be little hesitation about introducing it here. Frequently the term *valence-electron* is used instead of "loosest electron"; there is little in its favour beyond the general inability of physicists to think of a better one.[6]

Detachment of the Loosest Electron by Other Agencies than Electron-Impacts

Other agencies than the blows of electrons are capable of detaching the loosest electron from an atom; but it is very much more difficult to obtain simple and intelligible information about their immediate effects than about those of electron-impacts.

The study of *ionization by radiation* involves a host of new problems. Theoretically the conditions seem simple enough. Radiation of any frequency ν behaves in some respects as though it consisted of streams of particles each having energy $h\nu$ and momentum $h\nu/c$. Since it behaves in this manner in so far as absorption in gases and ejection of electrons from solids are concerned, we should expect it to do likewise in effecting ionization of atoms. If so, radiation should ionize atoms if and only if its frequency ν equals or exceeds a critical or threshold value ν_0, expressed in terms of the ionizing-potentials V_0 of the atoms (measured in equivalent volts) by

$$h\nu_o = eV_o/300. \tag{6}$$

Projecting light from a spectrum upon a gas, and passing steadily from low to high values of ν, we should expect ionization to commence abruptly at ν_0.

Experimentally, the task of testing this inference has baffled everyone, at least until very recently. In the first place, the values of threshold-frequency ν_0 for various atoms correspond to values of threshold-

[6] The terms "optical electrons" and "series electrons" are sometimes seen; they are derived from theoretical pictures which are in danger of mutation (some people now ascribe most series-spectra to displacements of electrons in groups). The German term "Leuchtelektron" probably sounds better in German than its equivalent "shining electron" would sound in English. It may be remembered that difficulty in choosing a good name for a concept sometimes signifies that the concept is essentially vague and not rooted in Nature.

length λ_0 lying between 504A (helium) and 3184A (caesium); and this is the most troublesome region of the spectrum to deal with, partly because light of wavelengths lying within it is tremendously absorbed by nearly all solids and even gases, and partly because good sources for such light are difficult or impossible to procure. Even in the comparatively accessible zone between 2000A and 3500A it is customary to use the light of the mercury arc, which provides a few widely-spaced bright spectrum-lines; as though in determining ionizing-potentials by electron-impacts one had to use electrons of certain distinct and widely-spaced energy-values, and could not refine the measurements by adjusting the accelerating voltage to intermediate values *ad libitum*. In measuring ionizing-potentials by electron-impacts there is a secondary difficulty due to radiation from struck atoms falling upon the collector; here the difficulty becomes a primary one, since the primary radiation itself is competent to produce this effect. The effect is most vicious with alkali-metal vapours, as they deposit themselves over all the solid surfaces of the apparatus in films excessively liable to pour out electrons when stimulated by light or warmth; yet these are the only elements for which λ_0 lies above 2500A.

Several experimenters have minimized the undesired effects of the radiation by projecting a narrow beam of light across a jet of alkali-metal vapor boiling up out of a narrow channel in the main tube. The beam struck nothing except the jet and beyond it a "trap" in which presumably it was totally absorbed and no part was scattered. The jet passed onward, near to an electrode negatively charged to receive positive ions. With potassium vapors, for which λ_0 should be 2856A, R. C. Williamson found ionization commencing somewhere between 3100A and 2800A; H. Samuel thought that it commences between 2804A and 2893A; E. Lawrence concluded that it begins at 2610A.[7] P. D. Foote and F. L. Mohler[8] detected the positive ions by their effect in annulling the space-charge limitations upon the current from a hot filament, after the fashion of the last-mentioned method of determining ionization-potentials. Their result was somewhat unexpected; they found ionization in caesium vapor at wavelengths even greater than the threshold-wavelength. This is attributed to the same cause as brings about a lowering of the apparent ionizing-potential when dense

[7] *Phys. Rev.* (2) 27, pp. 37–51 (1926); 26, pp. 197–207 (1925).

[8] E. O. Lawrence, *Phil. Mag.* 50, pp. 345–359 (1925); R. C. Williamson, *Phys. Rev.* 21, pp. 107 (1923); H. Samuel, *ZS. f. Phys.* 29, pp. 209–213 (1924); and prior literature cited in the first two. In all of the cited experiments the vapor had freshly issued from condensed potassium, and may have contained a large proportion of molecular aggregates, to which Lawrence attributes the difference between his observed threshold-wavelength and the calculated ν_0, Cf. also G. F. Rouse and G. W. Giddings, *Proc. Nat. Acad. Sci.* 11, pp. 514–177 (1925).

streams of bombarding electrons are used: that is to say, it occurs because light of less than the threshold frequency puts some of the atoms into abnormal states, in which less energy is required to ionize them than in the normal state.

The study of *ionization by positive ions* is also very troublesome. This is partly because there are no such convenient sources for controllable positive ions as there are for electrons. The ions emerging from hot filaments are generally not all of one kind. If ions of a particular sort, hydrogen ions for instance (these would give the most valuable information of any) are produced by bombarding the proper kind of gas by electrons having a suitable ionizing-energy, they cannot be used for ionizing except in the same tube and therefore upon the same gas; further, it is necessary to keep the bombarding electrons out of the region where the positive ions are meant to ionize, by an elaborate system of gauzes and opposing potentials. If the collecting electrode is maintained at a positive potential so as to receive electrons produced by the ionization, it receives also the electrons which are knocked out of the walls of the tube by positive ions which strike them. It is scarcely surprising, then, that the published data are scanty and not always concordant.[9]

The considerations about conservation of momentum during impacts, mentioned in dealing with ionization by electrons, show that we should hardly expect a positive ion to be able to ionize unless it has much more energy than must be transferred to the atom to detach the loosest electron from it; twice as much, if the ion is of the same mass as the atom.

Identification of Ions Produced by Electron-Impacts

The methods hitherto described for detecting the onset of ionization in a gas show when free positive charges appear in a gas, but give no further information about them. The methods employed by J. J. Thomson and F. W. Aston reveal the charge-to-mass ratios of ions occurring in a gas carrying a self-maintaining discharge, but give very little information about the precise conditions necessary to produce them. A combination of methods of these two kinds was first effected by H. D. Smyth.[10]

One of the tubes employed by Smyth is sketched in Fig. 3. Electrons from the filament F are accelerated through the potential-rise V_1 to the

[9] For work published up to 1922 see the review and bibliography by A. J. Saxton, *Phil. Mag.* 44, pp. 809–823 (1922). See also J. T. Tate, *Phys. Rev.* (2) 23, pp. 293–294 (1924).

[10] *Proc. Roy. Soc.* A102, pp. 283–293 (1922–23); A104, pp. 121–134 (1923); *Phys. Rev.* (2) 25, pp. 452–468 (1925) and references there given.

gauze E_1, and then turned back by an adverse potential-fall V_2 before they reach the partition E_2 pierced by the slit S_2. Positive ions produced by the electrons in the region between E_1 and E_2 are drawn

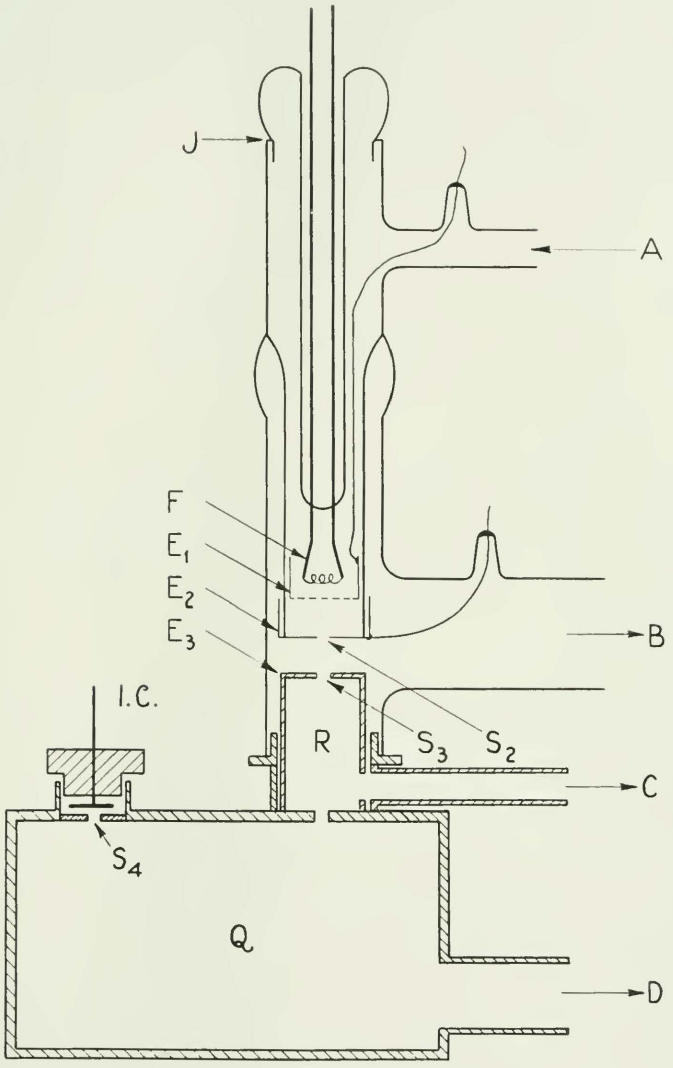


Fig. 3

toward E_2; some of them emerge through S_2, and encounter an additional potential-fall V_3 which draws them to the partition E_3. Those which pass through the slit S_3 are now ready, after passing through the

field-free region R, to be swung around in semi-circular arcs by a magnetic field H applied normally to the plane of the paper over the region Q; thus they arrive at the ion-collector behind the slit S_4. The major experimental difficulty consists in maintaining simultaneously a gas-density between F and E_2 high enough to afford plenty of ions, and a gas-density in R and Q low enough so that the ion-stream is not dispersed. This is effected by feeding in the gas through A and applying powerful pumps to draw it out through B, C and D.

Varying H and plotting against it the current into the ion-collector, one obtains a curve with peaks, such as the one in Fig. 4. This is, however, a curve obtained by Dempster with ions issuing from a hot filament. The charge-to-mass ratio for the kind of ion producing each peak is calculated from the accelerating-voltages, the deflecting field, and the diameter of the circular arc through which they swing.

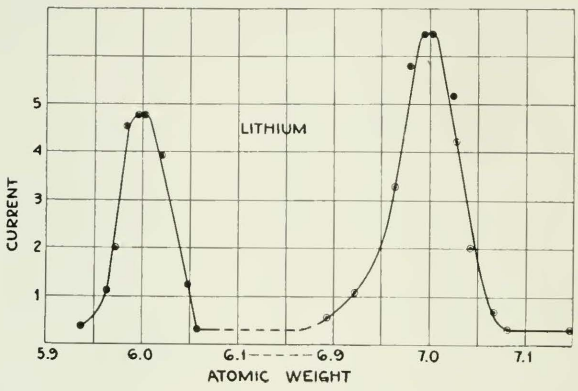


Fig. 4

The use of this method in determining ionizing-potentials may be illustrated from the work of H. A. Barton on argon.[11] Observing at values of V_1 superior to some 50 volts a two-peaked curve with the M/E values of the corresponding ions standing in the ratio 2:1; and observing at values of V_1 inferior to some 40 volts only one of these peaks, the one with the greater value of M/E; he inferred that this peak was due to A^+ ions and the other to A^{++} ions. Plotting the heights of these peaks or the areas under them as functions of V_1 he obtained curves such as those shown in Fig. 5. From many such curves as these he deduced that the energy of electrons just able to produce doubly-ionized argon atoms exceeds that of electrons just able to produce

[11] *Phys. Rev.* (2) 25, pp. 469–483 (1925).

singly-ionized argon atoms by 30 equivalent volts.[12] In the same manner, Smyth concluded that the energy of electrons just able to produce doubly-charged mercury ions exceeds by about 9 equivalent volts that of electrons just able to produce singly-charged mercury ions. The method, however, has been used chiefly for studying diatomic gases, and therefore will be mentioned in another section.

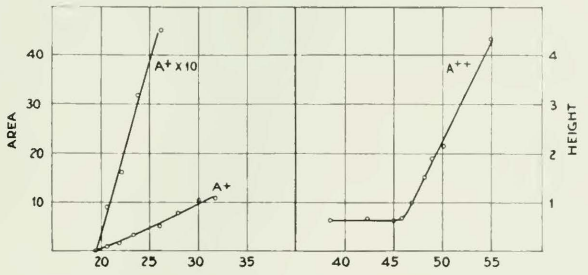


Fig. 5

Ionization of Molecular Gases [13]

The experiments of Thomson and Aston upon the ions proceeding from self-sustaining discharges in molecular gases show that these comprise individual atoms and also molecules of various sorts, each deprived of one or occasionally of more than one electron. Not all of these, however, are produced by the direct and simple agency of a single electron-impact against a normal molecule; some of them result from encounters of ions originally produced in the discharge with molecules which they meet in the gas, either in that region where the discharge is being maintained or in the channel through which they pass to reach the analyzing fields. This stands out very clearly in such experiments as one performed by A. J. Dempster, who projected 800-volt electrons into hydrogen gas and determined the relative abundance of the ions H^+, H_2^+ and H_3^+ arriving at his collecting-electrode after passing through a certain distance in the gas. At a gas-pressure amounting to .01 mm. Hg, the H_3^+ ion was the most plentiful of all and the other two not far behind; at .0017 mm. Hg both the H^+ and H_3^+ ions were definitely less abundant than H_2^+, and below .0005 mm. the H_2^+ ion

[12] Actually he obtained 17.3 volts for the one critical potential, 47.4 for the other, and assumed that the difference between 17.3 and the accepted value of 15.2 for the first ionizing-potential of argon is due to contact potentials and other influences affecting each of the observed critical potentials equally.

[13] For a general bibliography of this subject see T. R. Hogness & E. G. Lunn, *Phys. Rev.* (2) 26, pp. 44–55, 786–793 (1925); also V. Kondratjeff, *ZS. f. Phys.* 22, pp. 1–8 (1924) and 31, pp. 535–541 (1925).

was left almost alone upon the scene. These results signify that an 800-volt electron operates ionization in hydrogen by detaching an electron from a molecule; other kinds of ions appearing in the gas are due to subsequent adventures of these ions.

The method of H. D. Smyth is suitable for investigations into this question. In apparatus such as his, hydrogen bombarded by (say) 40-volt electrons is found to contain all three ions H^+, H_2^+ and H_3^+; but as the density of hydrogen is reduced, the first and the last of these ions become less abundant and finally insignificant by comparison with the ion H_2^+. As the bombarding-voltage is reduced towards the value (about 16) at which ionization commences, all three kinds of ions become less plentiful; but with high densities and sufficiently sensitive apparatus it is found that H_3^+ makes its appearance as early as H_2^+, and there is no reason not to suppose the same about H^+. In hydrogen, therefore, and also in nitrogen, it is agreed that an electron-impact against a molecule results, if in any sort of ionization at all, in the detachment of an electron from the molecule, not (for instance) in a dissociation into one ionized atom and another atom ionized or neutral. Dissociation and new sorts of association may result from the further adventures of this molecule-ion in the gas. In certain compound gases[14] of which the molecules consist of two or more atoms of different kinds, there is reason to expect the contrary: that is, that an electron-impact against a molecule would result directly in splitting it into a positively-charged atom (or group of atoms) and a negatively-charged atom (or group of atoms). Certain experiments indicate this: in $ZnCl_2$ vapor, for instance, Cl atoms bearing an extra electron and ZnCl molecules minus an electron are found as soon as ionization commences; but the question can hardly be deemed settled until comparative measurements are made at various gas-densities.

From these experiments it follows that a measurement of the energy just sufficient to produce ions in a molecular gas, while interesting in itself, can hardly be interpreted without additional data regarding the nature of the ions produced. There are other difficulties in determining ionizing-potentials in such gases; for instance the likelihood that the hot filament will itself dissociate the gas. The published determinations are frequently contradictory; the various published values for the ionizing-potentials of hydrogen, for instance, form one of the most discouraging sets of irreconcilable data to be found in physics.

According to thermochemical measurements the "heat of dissociation" of hydrogen, in other words the energy-difference between a system of two free H atoms and an H_2 molecule, amounts to 3.5 equivalent

[14] Those designated by chemists as heteropolar.

volts. One would expect to be able to dissociate hydrogen by bombarding the gas with 3.5 volt electrons; yet nothing of the sort happens. This is an instance of the frequently-occurring observation that a particle or a quantum may have abundant energy to produce a particular effect and yet be quite unable to produce it. One would expect also that the minimum energy required to convert an H_2 molecule into an H^+ ion and an H atom and a free electron would exceed by 3.5 equivalent volts the ionizing-energy of an H atom, yet the difference appears to be less, which is strange.

Detachment of Tightly-bound Electrons from Atoms

We will now consider the most direct and striking evidence for the statement that each atom (apart from those of the lightest elements) possesses several distinct ionizing energies—several distinct "states of ionization." This fact is taken to mean that each atom possesses several or many electrons which are *bound*, as the phrase is, with different degrees of firmness or tightness; that the ionizing-energies of the atom are, so to speak, the *extraction-energies* of these various electrons; to each electron there corresponds a certain extraction-energy, the amount of energy which must be imparted to the atom to extract that electron, the energy-difference between the normal state of the atom and that particular "state of ionization" which involves the absence of that particular electron. I shall frequently use the language of this interpretation, which is extremely convenient and likely to remain so. Nevertheless it is desirable to remember that the quantities actually observed are energy-differences between various states of the atom, or energy-values of various states of the atom referred to the energy-value of the normal state as zero. These energy-values are the data of experience; most other assertions about the states of ionization are speculative.[15]

Conceive a layer of atoms of an element possessing several different values of ionizing-energy W_1, W_2, W_3 and so forth; in other words, atoms which are capable of several states of ionization of which the energy-values exceed that of the normal state by W_1, W_2, W_3 and so forth. Suppose that a beam of radiation of frequency ν, so chosen that the product $h\nu$ exceeds all of the ionizing-energies, falls upon the layer. Such a beam is absorbed as though it consisted of individual particles of energy $h\nu$, each of which is either completely absorbed or totally ignored by the layer of matter upon which it falls. Consider an atom which ab-

[15] In some cases, although not in any which will be discussed in this section, it is found necessary to suppose that several distinct states of ionization correspond to the absence of a particular electron, which is somewhat of a strain upon the picture.

sorbs the amount $h\nu$ of energy from the beam. Through this absorption, an electron is detached from the atom. If however the electrons were merely separated from the atom and left stationary beside it, the energy of the system (ion plus electron) would by definition have been augmented merely by W_i. This quantity is (by our supposition) less than $h\nu$. However the entire energy $h\nu$ has been absorbed; the difference $(h\nu - W_i)$ is likewise transferred to the ion-plus-electron system, in the form of kinetic energy of the liberated electron. The electron flies away with speed V_i determined by the relation

$$\tfrac{1}{2}m V_i^2 = h\nu - W_i. \tag{7}$$

The foregoing paragraph contains several interlocking assumptions, which if they are all true lead to this conclusion: *When a beam of radiation of frequency ν falls upon a layer of atoms having ionizing-energies W_1, W_2, . . . W_i . . . , electrons of various speeds spring out of the layer,*

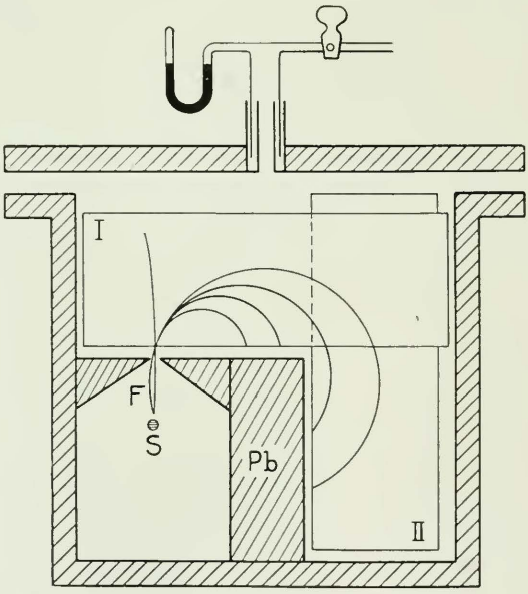


Fig. 6

there being for each value of W, a corresponding group of electrons of which the speed is given in terms of W by equation (7).

Suppose that one irradiates a metal with high-frequency radiation, and by a system of slits confines his experimentation to electrons projected in directions nearly normal to the metal surface, and applies a

magnetic field in a direction parallel to the surface. Then we have the situation which occurs in measuring the speeds and charge-to-mass ratios of electrons and ions by the method of electric acceleration followed by magnetic deflection. The only differences are, that in the present case the speeds v with which the electrons enter into the magnetic field are imparted to them not by an imposed electric field but by

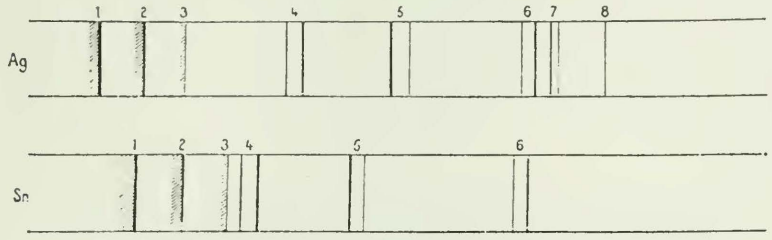


Fig. 7

the radiation which released them; and that the experimenter takes the value of e/m for granted and computes the values of v from the magnetic deflections alone. The electrons are swept around in circular arcs, of which the radii yield their speeds.

The apparatus by which such experiments are performed is of the type shown in Fig. 6. At S there is a long narrow rod or tube of the



Fig. 8

material to be tested; it is irradiated by X-rays proceeding from a source beyond the diagram to the left. A magnetic field, directed normally to the plane of the paper, sweeps the emerging electrons around in circular arcs, some of which pass through the slit. The appearance of films laid along the top of the block Pb, normal to the plane of the paper, is shown by Figs. 8 and 9. They suggest spectra; and though the lines are signatures of special electron-speeds rather than of special

radiation-frequencies, the difference between these is not so radical as once it seemed, and we may without hesitation call them by some such name as *electronic spectra*.

Each line in such a spectrum is produced by electrons of a definite extraction-energy, extracted by radiation of a definite frequency. Continuing with the policy of referring to electrons with a definite extraction-energy as being definitely individualized within the atom, I will designate the electrons of greatest extraction-energy for any particular kind of atom as the K electrons; those of next greatest extraction-energy as the L electrons, and then the M and N electrons in due order. (Later it will be necessary to subdivide these classes, but for the moment this may be avoided.) At other times I shall speak of these elec-



Fig. 9

trons as belonging to the K level, the L level, and so forth; still other terms in use are the K shell and the L shell, or the K ring and the L ring. In a given electronic spectrum we may expect to find a set of lines due to K, L, M and other electrons, for each frequency represented in the incident radiation; unless there are some of these frequencies for which the quantum energy $h\nu$ is less than the extraction-energies of some of the electron-groups, in which case there will be no corresponding lines.

An ideally simple electronic spectrum would be produced by a single radiation-frequency; but this is impracticable, for even if one were to eliminate from the stream of X-rays proceeding out of an X-ray tube all but one frequency, the irradiated atoms would themselves supply others.[16] As in the mass-spectra upon Aston's plates, this unavoidable complexity is actually an advantage; it helps in identifying the several lines.

In Figs. 8 and 9, photographs taken in the manner already mentioned, there appears the electronic spectrum due to silver atoms irra-

[16] These are in fact especially efficient in ejecting electrons, as they originate within the atom-layer itself.

diated by the characteristic X-rays of tungsten.[17] To guard against the possibility that the photographs may lose in clearness by the process of reproduction, I will base the explanation upon the uppermost of the sketches in Fig. 7, which is abstracted by de Broglie from similar pictures. The electron-speeds corresponding to the lines increase from left to right. The irradiating X-rays consist of four characteristic frequencies from the X-ray spectrum of tungsten; in order of decreasing frequency they are known as $K\gamma$, $K\beta$, and the two members of the $K\alpha$ doublet. The four lines marked 4 and 5 in the electronic spectrum are made by electrons extracted by these four radiations from a single level —the K level of the silver atoms. The two following doublets, marked 6 and 7, are made by electrons extracted by the $K\alpha$ frequencies from two other levels of the silver atom, the L and M levels respectively. Line 8 is due to $K\beta$ extracting electrons from the L level. At the other end of the spectrum, the three lines 1, 2, 3 are due to electrons ejected from the L and the M levels by two of the X-ray frequencies characteristic of silver, which the irradiating X-rays stimulate some of the silver atoms to emit. The rays responsible for these particular lines are the so-called $K\alpha$ and $K\beta$ rays of silver, which are so related to one another (as will be stressed in a later passage) that the electrons extracted by the former from the M level have very nearly the same energy as the electrons extracted by the latter from the L level, so that the two frequencies acting on the two groups of electrons produce three (instead of four) distinct lines of the electronic spectrum.

Reverting now to the photographs: in Fig. 8 the pairs of lines marked 4, 3, and 2 are those designated respectively as 6, 5 and 4 in the sketch and in the foregoing explanation, while the lines to the left are those produced by characteristic X-rays of silver acting upon silver atoms. On a larger scale, this latter region of the spectrum is shown in Fig. 9; here the lines are marked by the same numerals as in the sketch; the pair at 4 is due to K-electrons extracted by the two $K\alpha$ rays of tungsten, the line 3 is due to M-electrons extracted by the $K\beta$ radiation of silver, the line 2 results jointly from L-electrons extracted by the $K\beta$ radiation of silver and M-electrons expelled by the $K\alpha$-radiation of silver, while the line 1 is due to L-electrons ejected by the $K\alpha$-rays of silver.

The resemblance and the differences between electronic spectra of elements not far apart in the Procession are illustrated by the two sketches in Fig. 7, the lower relating to tin (atomic number 50) and the upper to silver (atomic number 47) irradiated by the same frequencies.

[17] I am greatly indebted to M. de Broglie for sending me the negatives of these admirable pictures, as well as that of Fig. 10.

Since the extraction-energy of each named class of electrons increases along the periodic table, the lines designated as 4, 5 and 6 in the electronic spectrum of silver reappear in that of tin, displaced in the direction of diminishing electron-speeds, that is, to the left. But, as to the lines 1, 2, and 3, both the extraction-energies of the electrons and the frequencies of the rays responsible for these alter as one passes from silver to tin, and the net result of the double alteration is that the lines are displaced to the right.

The energy-values of the various states of ionization of an atom—or, in terms of the customary picture, the extraction-energies of the various classes of electrons within the atom,—may be determined with a certain degree of precision from experiments such as these. However, as in the case of the measurement of charge-to-mass ratios for individual ions by the methods of Aston and Dempster, there is little incentive to develop the accuracy of the method to the highest possible extent; for most of the energy-values in question can be determined with very great accuracy in another way, which we will now examine.

Absorption of Radiation Through Ionization

When a beam of radiation of frequency ν is transmitted through a layer of matter, from the atoms of which it extracts electrons with an expenditure of energy $h\nu$ at each extraction, we should expect to find it correspondingly reduced in intensity when it emerges from the layer.

This effect is strikingly conspicuous with radiation high enough in frequency to detach the tightly-bound electrons of massive atoms. Let a narrow beam of "heterogeneous" radiation, containing all frequencies throughout the widest possible range, fall from an X-ray tube through slits and diaphragms upon a thin layer of such atoms; let the transmitted rays be dispersed by some appropriate spectroscope, and fall finally upon a photographic plate on which their spectrum—in the ordinary sense of the word, not in the sense of "electronic spectrum"—is outspread.

In Fig. 10 there are three such spectra, of heterogeneous beams which have passed through layers of cadmium, antimony, and barium respectively. The frequency increases from right to left. The darkening at any point is a measure of the intensity with which the X-rays acted at that point.

Below a certain frequency identical for all three elements, the photographic films have evidently been little affected; as soon as this critical frequency is exceeded, the effect suddenly becomes enormous. This critical frequency is the one for which the quantum-energy just

suffices to extract a K-electron from a silver atom; for the photographic film contains silver, and it is the expulsion of electrons from the atoms in it which initiates the photographic process. Proceeding always toward higher frequencies, we see that presently the plates suddenly become whiter, at another critical frequency which however



Fig. 10

is not the same for the three elements. The photographic film is not responsible for these "absorption-edges" as they are called; each of them occurs at the particular frequency for which the quantum-energy just suffices to extract a K-electron from an atom of the element which formed the absorbing-layer placed in the path of the beam before it reached the plate. To the right of the absorption-edge we have the lower frequencies, unimpeded by the cadmium (or antimony, or barium) atoms because unable to ionize them; to the left we have the higher frequencies, reduced in intensity by the intercalated matter because some of their energy was drawn off to detach electrons.

From the frequency ν at such an absorption-edge, the extraction-energy W of the class of electrons in question for the kind of atom in question is determined by the equation

$$h\nu = W.$$

This is a much more delicate way of measuring extraction-energies than the observations upon electronic spectra afford. Nevertheless the measurements upon the energies of the ejected electrons are of the greatest importance, for they show what is effected by the energy-transformations which set in when one or another of these critical frequencies is overpassed.

Likelihood of Ionization by Electrons Having More Than the Least Ionizing-Energy

We have seen that electrons projected into a gas of ionizing-energy V_0 are able to ionize it if their kinetic energy exceeds V_0, otherwise not (apart from ionizations effected upon atoms in abnormal states). This

question now suggests itself: Suppose that a great number Q of electrons, all having kinetic energy V, falls upon a thin stratum of gas containing dN atoms per unit area: how many atoms will they ionize, how many ions will be produced? Designating this number by $Qf(V)dN$: what is $f(V)$?

This question is much easier to formulate in words than to answer by experiment. Suppose for instance that one should try to answer it by means of the scheme of apparatus sketched in Fig. 1. In going from G_1 to G_2, coming to a stop between G_2 and G_3, and returning again, the electrons pass successively through all values of kinetic energy from their highest down to zero and back to their highest again; and ions are produced by electrons of all values of kinetic energy, from their highest down to the ionizing-energy. The ions collected by the collector at C represent a sort of integral of $Qf(V)dN$ taken between V_o as minimum and the energy possessed by the electrons at G_1 as maximum. To determine $Qf(V)dN$ it is necessary to measure the total ionization at several values of V and then construct a sort of differential curve. To determine Q it is necessary to know how many electrons come from the filament into the ionizing-region, and in addition how many extra ones are introduced through primary electrons knocking them out of the gauze of G_1.

Another scheme consists essentially in making G_2 into a solid wall and using it to collect the electrons, so that after passing from G_1 to G_2 they vanish from the scene. This would be excellent if the region between G_1 and G_2 could be left equipotential; but it is necessary to intrude a negatively-charged electrode in order to collect the positive ions, and apparently whenever this electrode is sufficiently large and sufficiently negative to capture the ions it is also sufficiently large and sufficiently negative to distort the field between G_1 and G_2 quite seriously; so that the electrons are at first slowed down and later speeded up again as they pass from G_1 to G_2, and the ions received by the collector are as before a sort of integral of $Qf(V)dN$.

In spite of these difficulties the various experiments performed with extremely rarefied gases yield fairly concordant results.[18] The function $f(V)$ mounts steadily, from zero at the ionizing-energy V_o, to a broad and flattish peak culminating somewhere between 100 and 400 volts (depending on the gas), and thereafter declines slowly as V increases. Thus, although an electron striking an atom (or molecule) can detach the loosest electron if it has just the requisite energy, its chance of doing so is improved if its energy is greater than the just-sufficient

[18] K. T. Compton and C. C. van Voorhis, *Phys. Rev.* (2) 26, pp. 436–453 (1925) and literature there cited; also W. P. Jesse, ibid. pp. 208–220.

amount. However, it would not be safe to infer that throughout the range of these observations all of the ionizations consist in detachments of valence-electrons from various atoms. Sooner or later transfers of atoms into other states of ionization must commence. This is rendered all the more probable by the fact that the values of $f(V)$, determined at or near the peak for each gas, show a very definite tendency to increase steadily with the number of electrons in the atom or the molecule in question.

If a stream of electrons is projected into a sufficiently dense gas, the electrons are gradually slowed down and even stopped, and the stream is dispersed. Measurements of the number of ions produced per electron per millimetre have been made under such conditions,and measurements also of the "total ionization" produced in a volume of gas so large that the electrons lose their forward speed altogether before reaching the walls; but though the intrinsic interest of such measurements is great, it seems practically impossible to deduce $f(V)$ from them.[19] The difficulties may be compared with those arising in the study of alpha-particle scattering when the metal foil is too thick.[20] When, however, the electrons are moving with the enormous speeds possessed by those ejected from radio-active substances, or when ionization by alpha-particles is studied, the conditions again become simpler and relatively intelligible.

Ionization by Alpha-Particles and Very Fast Electrons

Ionization by particles possessing kinetic energies amounting to millions of equivalent volts, such as alpha-particles and many of the electrons emerging from radioactive substances, might well be expected to follow other laws than ionization by particles possessing little more than enough energy to detach an electron from an atom. Such indeed is the case; yet it would not be justified, either by reasoning or by experiment, to suppose that even such highly energetic particles expel electrons of any and every class tightly-bound alike and loosely-bound alike, with equal ease and abundance from the atoms which they strike.

It is not particularly difficult to measure the total number of ions produced by an alpha-particle in its course through a gas from the moment it enters, with a measurable initial speed, to the moment when it goes into retirement (so to speak) as an ordinary helium atom; nor to

[19] G. A. Anslow, *Phys. Rev.* (2) 25, pp. 484–500 (1925) and literature there cited.

[20] Anyone desiring to learn how complicated the circumstances may become when electrons are shot into a dense gas should read P. Lenard's brochure "Quantitatives über Kathodenstrahlen," published by the Heidelberg Academy in 1918.

divide this number into the kinetic energy which it originally had, thus obtaining the average energy spent per ion (or rather per pair of ions generated, since each ionization produced two ions of opposite sign)—a quantity amounting generally to several tens of equivalent volts (33 volts for air).[21] By performing such experiments with alpha-particles of various initial speeds, it is possible to determine a function analogous to the function $f(V)$ defined for electrons in a previous section. This function increases rapidly as the aplha-particle approaches the end of its sharply-terminated trail, varying approximately as the reciprocal of the cube root of the distance it has yet to go.

Alpha-particles as they pass through a gas thus produce a countable number of ions and suffer a measurable loss in kinetic energy. It is interesting to enquire whether these two processes can be identified with one another and explained by the nuclear atom-model—whether the lost kinetic energy is altogether spent in detaching electrons from the atoms of the gas and supplying them with extra kinetic energy.

Before comparing any theory with the experimental data, one must be aware of two complexities. In the first place, an alpha-particle may transfer energy to an atom without ionizing it, so that the energy it loses in passing through a gas may exceed that which it spends in ionizing. In the second place, some of the ions produced by an alpha-particle—notably, the detached electrons—may themselves be endowed with energy enough to ionize, so that a measurement of the total ionization in the gas may yield an excessive estimate of the number of ions actually and immediately produced by particles striking atoms. Naturally the energy for producing all of these ions, "primary" and "secondary" alike, comes from the alpha-particles, so that such data as the aforesaid values for energy-spent-per-ion-generated have a definite meaning.

Discrimination between ions produced directly and indirectly is desirable, indeed essential, for testing any theory; but thus far there is no way for distinguishing the two, except in the case of very fast electrons for which the trails have been photographed with great magnification by C. T. R. Wilson[22] by his celebrated expansion-method, in which each ion formed in the passage of such an electron through a gas becomes the center of a visible droplet of water. In some of his pictures, in Fig. 11, pairs of droplets and also groups of four, six and more are seen. The paired droplets have condensed upon the two ions, positive and negative, produced by a single primary ionization (and

[21] R. W. Gurney, *Proc. Roy. Soc.* A107, pp. 332–340 (1925) and literature there cited.

[22] *Proc. Roy. Soc.* A104, pp. 192–212 (1923).

then drawn apart by an appropriate electric field); the groups of more than two bear witness of a primary ionization followed by secondary processes of the same type. In Fig. 12 there is an actual long branch to the primary trail; the original fast electron has detached another and endowed it with so great an energy that in ionizing-efficiency it



Fig. 11

rivals its liberator. These figures show that a mere count of all the ions formed by a particle flying through a gas is no estimate of the detachments of electrons from atoms which the particle of itself and at first hand effected.

Various theoretical expressions have been derived for the rate of slowing-down and the rate of ionization of an alpha-particle or fast

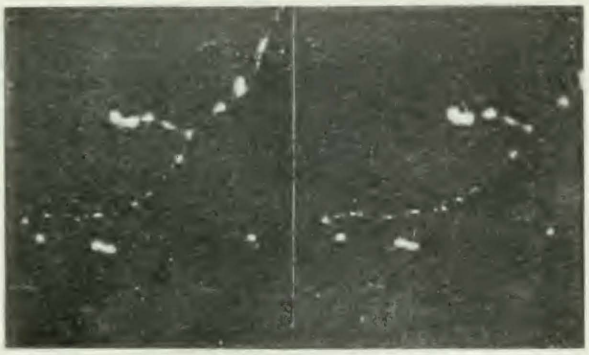


Fig. 12

electron proceeding through a gas. Most of them lead to what are known as "order-of-magnitude agreements," but none to a close quantitative agreement—which is, perhaps, after all better than could be expected. They are founded upon an equation originally proposed by J. J. Thomson. Suppose a stratum of an element of atomic number Z, containing N atoms; using the nuclear atom-model, we conceive this as a region containing N nuclei and NZ electrons. If the electrons (of

mass m) were free and stationary, an alpha-particle of mass M moving with speed U along a line passing at distance p from the initial position of any one of them would communicate to it an amount of energy:

$$W = \frac{8e^2}{mU^2(p^2+a^2)}$$

where

$$a = \frac{2e^2(M+m)}{mUM^2}. \tag{8}$$

Imagine Q alpha-particles passing through this collection of NZ electrons; the number of encounters for which this energy-value lies between two values W and $W+dW$ is equal to $2\pi p(dp/dW)dW$. Multiplying this by W and integrating over all values from $W=0$ (corresponding to $p=\infty$) to $W=8e^2/mU^2a^2$ (corresponding to $p=0$), we arrive at a value for the total amount of energy communicated by the alpha-particles to the electrons, which value is infinite. This absurd conclusion rests on the absurd assumption that the electrons are free, which, of course, is not made. Generally it is assumed that whenever p exceeds a certain value, selected for one reason or another, equation (8) loses its validity and W is zero; for instance, that whenever p is so great that the value computed by (8) for W is smaller than the least energy sufficing to remove the electron altogether from the atom or to put the atom into a Stationary State, then there is no transfer of energy whatever; but, whenever p is so small that W as computed by (8) exceeds the extraction-energy for the electron in question, then the electron is extracted and carries off, as kinetic energy, the difference between W and its extraction-energy.

Definite assumptions must be made about the extraction-energies of the various classes of electrons in the atom, the number of electrons in each class, and the Stationary States of the atom; this being done, formulae are derived for the primary ionization, the scondary ionization, and the rate at which the alpha-particle (or fast electron) loses energy.[23] Apart from these results of elaborate and careful analysis which lead as I have said to order-of-magnitude agreements (in some cases the agreements approach quantitative value) it may be pointed out that the equation (8) leads, when U is so great that a becomes small relatively to p, to the conclusion that as a fast-flying particle proceeds through matter the fourth power of its speed falls off linearly with increase of distance traversed, which is in agreement with much

[23] See R. H. Fowler, *Proc. Camb. Phil. Soc.* 21, pp. 521–540 (1923), and G. H. Henderson, *Phil. Mag.* 44, pp. 680 (1922) for discussion and prior literature as well for their own work.

experimental work. It furthermore indicates that the total ionization effected by a beam of particles in traversing a given thickness of matter should, beyond a certain speed, diminish with increasing speed; which for alpha-particles is true for the entire available speed-range, and for electrons is true beyond the speed of optimum ionizing-efficiency mentioned in a previous section.[24]

Multiple Ionization

The analyses of positive rays issuing from gases sustaining electrical discharges show that under such conditions some atoms are deprived of two, three, or even so many as eight electrons. The recently-developed methods of interpreting spectra make it practically certain that some of the spectrum-lines emitted from gases bombarded by electrons or sustaining discharges, and particularly from the exceptionally violent discharges known as "sparks," are due to atoms lacking one, two, or so many as six of their normal complement of electrons.[25]

Such ions might conceivably be produced either in one operation or in several; that is, the two (or more) missing electrons might have been removed by a single agency at a single moment, or they might have been detached one after the other by separately and successively acting agents. Measurements by the method of H. D. Smyth, such as those upon argon already cited, are capable of showing the minimum amount of energy which bombarding electrons must possess, in order that doubly-ionized atoms may appear in a bombarded gas; but they do not show, at least not directly, whether this minimum amount is what is required to effect double ionization in a single operation, or merely what is required to effect the most difficult among two or several steps leading cumulatively to the result. The same holds true about the experiments in which the least bombarding-voltage sufficient to bring out the spectrum associated with the doubly-ionized atom is measured.[26] Granted that the energy-difference between the once-ionized and the normal argon atom is 15 equivalent volts, and

[24] The attempts to account for "straggling" of alpha-particles—that is, for the fact that different particles of the same initial speed are slowed down at somewhat different rates in progressing through the same gas—by ascribing it to mere statistical fluctuations in the number of electrons close to which they passed seem to have been unsuccessful; the observed straggling is much too great for this explanation. See G. H. Henderson, *Phil. Mag.* 44.

[25] Multiply-ionized atoms are regularly observed in electrolytic solutions of compounds of other-than-monovalent elements; strangely enough they are rarely if ever found among the ions issuing spontaneously from hot metals and salts.

[26] P. D. Foote et al., *Phil. Mag.* 42, pp. 1002–1015 (1921); *Astroph. Jour.* 55, pp. 145–161 (1922); *Origin of Spectra*, 1922.

that A^{++} ions appear in argon bombarded by 45-volt electrons: do 45 equivalent volts constitute the amount of energy necessary to remove two electrons at once from a normal atom, or the amount necessary to remove one electron from an atom which a prior electron-impact has ionized? The question is not different in principle from one arising in measurements of the first ionizing-potential, whether the first appearance of ions signifies simply that atoms are being ionized in two stages; but apparently it is harder to settle by direct evidence.

Analysis of the spectra of the ion and of the atom whenever practicable, discloses definitely the energy-differences between the state of double ionization, the state of single ionization, and the normal state of the neutral atom. Thus with helium the first of these is greater than the second by 54 and then the third by 79 equivalent volts. Similar calculations for magnesium show that the first is greater than the second by 15 equivalent volts; as the spectrum of the ion Mg^{+} in Foote's just-cited experiments appeared at about that energy of the bombarding electrons, the atoms in the vapor must have been ionized by two successive impacts.

In the course of R. A. Millikan's observations upon droplets of oil floating in ionized gases, he found that they never captured charges amounting to $2e$ or a greater multiple of e, except in the solitary instance of helium traversed by alpha-particles; in this case about one out of every six positive charges captured was a double electron-charge $2e$. He concluded that not more than one electron was ever detached from an atom in a single operation, except that among encounters of alpha-particles with helium atoms about one-sixth caused both of the electrons of the struck atom to be torn away.[27]

Detachments of two or more tightly-bound electrons from a massive atom, whether effected in one operation or in several, might be revealed by additional absorption-edges in the spectrum of an X-ray beam after passing through matter; certain delicate features in X-ray spectra have in fact been explained in this manner.

Thermal Ionization [28]

In addition to all the information about ionization by particular agents such as electrons of specified speeds and radiation of specified frequencies, there is reason for making certain assertions about ioniza-

[27] The percentage may well have been much greater, since many of the ions left behind after the passage of the alpha-particle were probably produced by secondary, not primary ionization (R. H. Fowler).

[28] General references: E. A. Milne, *Proc. Phys. Soc. London* 36, pp. 94–113, and literature there cited; A. A. Noyes, H. A. Wilson, *Pro. Nat. Acad. Sci.* 8, pp. 303–307 (1922).

tion *per se*, apart from all knowledge or assumption concerning the processes which effect it. There is a thermodynamic method of determining the percentage of dissociated molecules in a molecular gas as a function of the temperature and the pressure of the gas, which can be used if we know the amount of energy required to dissociate a single molecule, the specific heats of the undissociated and the dissociated gas, and the chemical constants of the undissociated and the dissociated gas. An analogy may be established between dissociation and ionization: the ionizing-energy of a monatomic gas corresponds to the heat of dissociation of (say) a diatomic gas; the electrons and the ions resulting from the ionizations may be taken as the particles of two distinct gases mingled with one another and with the gas composed of the neutral atoms; the chemical constant of the ion-gas is taken as equal to the chemical constant of the original gas, and the chemical constant of the electron-gas is identified with that which a gas composed of neutral atoms, each possessing the same mass as an electron, would possess. Utilizing this analogy, a formula may be deduced for the percentage of ionized atoms present in a monatomic gas in thermal equilibrium at any temperature and pressure.

Without developing the formula, it may be taken as a rather obvious inference that the higher the temperature of the gas at a given pressure, or the lower the pressure at a given temperature, the greater the percentage of ionization will be; and of two gases maintained at the same temperature and pressure, the gas having the smaller ionizing-energy will be the more ionized.

Measurements of the degree of ionization in a flame of known temperature, into which a known amount of caesium was introduced, have yielded values in good agreement with the percentage calculated from the thermodynamic formula; and measurements upon the conductivities of the vapors of the alkali metals have shown that they stand in the order of the ionizing energies reversed, although in other respects the agreement with the theory is not good.[29]. The tests and the value of the theory, however, appear chiefly in the realm of astrophysics. The hotter the region of a star in which the lines observed in its spectrum have their source, the more the lines of ionized atoms predominate among these. In many cases it happens that lines of ionized atoms are the only ones characteristic of a given element to be found at all. The assertion once commonly made, that certain elements are absent from the sun or other stars, is invalidated by the fact that under the actual conditions of temperature and pressure prevailing in these bodies,

[29] B. T. Barnes, Phys. Rev. (2) 23, pp. 178–188 (1924); M. N. Saha, Phil. Mag. 46, pp. 534–543 (1923).

those elements if present would be totally ionized and would not reveal their familiar lines at all. Rubidium was thought to be omitted from the composition of the sun, until it occurred to H. N. Russell to look for its lines in the spectra of comparatively cool sunspots. The relative intensities of the lines of ionized and non-ionized atoms of various kinds in the spectra of individual stars are now ascertained and used as a guide in assigning temperatures to these stars, and their guidance is shown reliable by the accord between the conclusions to which it leads and conclusions otherwise attained. The study of ionization in the laboratory thus contributes to the understanding of the stars.

Methods of High Quality Recording and Reproducing of Music and Speech Based on Telephone Research[1]

By J. P. MAXFIELD and H. C. HARRISON

Synopsis: This paper deals with an analysis of the general requirements of recording and reproducing sound without appreciable distortion. The storing or recording of sound requires, first, a mechanical system which will respond faithfully to the sound waves which are to be recorded. Then there is required some material in or on which this sound may be recorded and an intervening system which permits the sound waves to make the record in this material. In the usual case, and in that which is particularly discussed, there is a mechanical system which will vibrate in response to the sound which is to be recorded and directly through some mechanical linkage, or less directly through an electrical linkage, drives a cutting mechanism which will impress a wax record.

The amount of power available to operate the recorder directly from the sound in the recording room is so small as to make the use of high quality electrical apparatus with associated vacuum tube amplifiers of very distinct advantage over the acoustic method.

Where the question of reproduction is concerned, the same two alternatives mentioned for recording present themselves, namely, direct use of power derived from the record itself vs. the use of electro-mechanical equipment with an amplifier. In this case, however, the situation is materially different since the power which can be drawn directly from the record is more than sufficient for many uses. It is, therefore, generally simpler to design one single mechanical transmission system than it is to add the unnecessary complications of amplifiers, power supply and associated circuits. In cases where music is to be reproduced in large auditoriums, the power which can be drawn from the record may be insufficient and some form of electrical reproduction using amplifiers becomes necessary.

The paper points out, at length, how many of the heretofore unsolved fundamental problems of sound recording and reproduction have been readily solved by the application of a detailed knowledge of telephone transmission theory. The advances which have been effected in telephone transmission theory and in related electrical measuring apparatus in the last few years, have been so great as to surpass previous knowledge of mechanical wave transmission systems. The result is, therefore, that mechanical transmission systems of the type here considered, and perhaps other types, can be designed more successfully if they are viewed as the analogs of electric circuits. A detailed analysis is here made of the analogies between electrical and mechanical systems in the voice frequency range and a discussion of the resulting mechanical design is presented.

Introduction

THE problem with which this paper is concerned, in its broadest sense, may be stated as that of taking sound from the air, storing it in some permanent way and reproducing it again without appreciable distortion. It is immaterial from the general standpoint whether the means used are mechanical or electrical or a combination of the two. The choice of which method to use will depend largely upon the commercial requirements accompanying the specific purpose for which the reproduction is being made. For instance, it is quite probable that

[1] As printed here this paper is essentially as read before the A.I.E.E. Feb. 8-11, 1926.

the means chosen for reproduction in residences would differ materially from those used in large ballrooms or in the presentation of synchronized motion pictures.

Before considering the methods and results referred to in the title of this paper, it may be well to make a rough division of the problem. The storing or recording of sound requires, first, a mechanical system which will respond faithfully to the sound waves which are to be recorded. Then, there is required some material in or on which this sound may be recorded and an intervening system which permits the sound waves to make the record in this material. In the usual case, and in that with which we are particularly concerned here, there is a mechanical system which will vibrate in response to the sound which is to be recorded and directly through some mechanical linkage or less directly through an electrical linkage, drive a cutting mechanism which will impress a wax record.

The first consideration, therefore, is the character of the sound which is to be recorded including all of the effects of reverberation and the general questions of studio design. Next to be considered is the manner in which the cutting instrument shall impress this speech or musical record upon the constantly rotating wax disk, which disk is commonly called the wax master. In this connection, there will be discussed also the relative value of the electrical and mechanical linking of the cutting knife with the mechanism which receives the sound waves. Following the discussion of these problems and a brief reference to the state of the prior art, there remains to be considered the reproduction of the sound which is stored in the cuts or grooves of the wax record.

In the case of reproduction also, there is required a mechanical system which will respond to these cuts in the wax and a system which will set up in the air-sound waves essentially identical to those picked up by the first mechanism of the recording system. Between these two systems, a mechanical linkage intervenes in the case under discussion, but reference is made to the relative advantages of this system compared with the use of an electrical linkage.

First to be described, is the character of the sound which is to be recorded and reproduced and the effects of reverberation and transients upon the listener's sensation of this sound.

Studio Characteristics and Transients

Phonographic reproduction may be termed perfect when the components of the reproduced sound reaching the ears of the actual listener have the same relative intensity and phase relation as the sound reach-

ing the ears of an imaginary listener to the original performance would have had. Obviously, it is very difficult, if not impossible, to fulfill all of these requirements with a single channel system, that is, with a system which does not have a separate path to each ear of the listener from the sound source.

The use of two ears, that is, two-channel listening, gives the listener a sense of direction for each of the various sources of sound to which at a given moment he may be listening, and, therefore, he apprehends them in their relative distribution in space. It has been found possible with a single channel system, however, by controlling the acoustic properties of the room in which the sound is being recorded, to simulate to a considerable degree in the reproduced music the effective space relationships of the original. In this case, with a one-channel system, the directional effect is, of course, entirely absent, and the spatial relationship which is apprehended is probably due to the increased apparent reverberation of the instruments situated at the far end of the room as compared with those in the near foreground.

In recording work, therefore, one of the important acoustic characteristics of a room is its time of reverberation. Although it is probable that this is the most comprehensive single factor, experiment has shown that the shape of the room and the distribution and character of the damping surfaces play a part in the excellence of music in such a room.

It has been shown by Sabine[2] that for piano music, studios should have a time of reverberation measured by his method of 1.08 seconds. Experience has indicated that this figure is also very closely correct for other types of music. This figure of Sabine's assumes binaural listening. With single-channel systems, such as most of the present reproduction systems, whether for radio or the phonograph, the ability of the listener to separate the reverberation from the direct music by means of the sense of direction is completely removed and there is thrust upon his attention an apparently excessive amount of room echo. Experiment has shown that a time of reverberation for the recording room ranging from slightly more than $\frac{1}{2}$ to slightly less than $\frac{3}{4}$ of Sabine's figure affords in the reproduced music the effect of a room with proper acoustics. When this effect is accomplished, the person listening to the reproduced music has the consciousness of the music being played in a continuation of the same room in which he is listening and also has a sense of spatial depth.

Experiment has indicated further that any transients set up by the recording or reproducing system constitute a second cause of apparent

[2] Collected papers of W. Sabine.

increased reverberation. The data obtained thus far are insufficient to permit assignment of quantitative values to the importance of these two factors.

At the present state of the art, the most important requirement of a recording or reproducing system is its frequency characteristic. This involves two factors—intensity versus frequency, and phase distortion versus frequency. The effect of the second of these factors is not thoroughly understood but as it is closely related to the production of transients it has to be considered, as mentoined above. The system to be described is, however, relatively free from violent phase shifts within most of the range covered, but does have some undesirable phase-shift characteristics with small accompanying transients near its limiting cut-off frequencies.

Frequency Requirements

The frequency range which it would be desirable to cover if, it were possible, with relatively uniform intensity for the transmission of speech and all types of music including pipe organ is from about 16 cycles per second to approximately 10,000.

It may be interesting to examine the record requirements for a band of frequencies this great. For the purpose of this illustration, a lateral cut record will be assumed although in all the factors except the time which the record will run, the arguments apply in a similar manner to the hill-and-dale cut. Since, for mechanical reproduction, the sound at a given pitch is radiated by means of a fixed radiation resistance, it is necessary that the record must be cut with a device the square of whose velocity is proportional to the sound power. Under these conditions, it is seen that for a given intensity of sound the amplitude is inversely proportional to the frequency of the tone, and that a point will be reached somewhere at the low end of the sound spectrum where this amplitude will be great enough to cut from one groove into the adjacent groove, or in case of vertical cut, to cut so deeply that with present materials the wax will tear instead of cut away with a clean surface. This means that there is an inherent maximum amplitude beyond which it is not commercially feasible to go. Similarly the minimum radius of curvature of sine waves of various frequencies cut at constant velocity is inversely proportional to the frequency, so that as higher and higher frequencies are reached the radius of curvature becomes smaller and smaller until finally it becomes too small for the reproducing needle to follow. There is, therefore, an inherent limit at the upper end.

In order to extend these limits, it is necessary in the case of the low

end to make the spiral coarser and in the case of the high end to run the record at a higher speed. Both of these changes tend to decrease the time which a record of a given size can be made to play. The only alternative of these methods is to cut the record less loud than is the present standard practise and make the reproducing equipment more sensitive. This could easily be done if it were not for the "record noise" or "surface noise," as it is commonly called. Since this surface noise is already loud enough in comparison with the reproduced music to be somewhat objectionable, no appreciable gain in this direction can be made until the technique of record manufacture has been distinctly improved.

In this connection, there is one other interesting point. It has been suggested that if electric reproduction were used, it would be possible to cut the record with a characteristic other than uniform velocity sensitiveness and correct for the error by an electrical system whose characteristic is the inverse of the characteristic of record. If the change which is made in the recording characteristic tends toward cutting at uniform acceleration sensitiveness, the amplitude varies inversely as the square of the frequency and hence the difficulties at the low end of the scale are greatly enhanced. Similarly, if the records are cut more nearly at constant amplitude, the radius of curvature of the sine waves decreases as the square of the frequency, hence the difficulties are placed at the upper end. In the process which is being described in this paper, these limitations have been met commercially by having a frequency characteristic of the uniform velocity type between the frequencies of 200 and approximately 4000 cycles per second. Below 200 it has been necessary to operate at approximately constant amplitude with a resulting loss in intensity which loss increases as the frequency decreases. Above 4000 it has been necessary to operate at approximately constant acceleration with its consequent slight loss in intensity at the very high overtones. With a characteristic of this type, a range of frequencies from 60 cycles to 6000 can be recorded with reasonable success although the very low and very high range are slightly deficient. (See Fig. 14) With a record having such a frequency characteristic, the inherent limitations are divided between the two ends of the frequency band and where electrical reproduction methods are used, it is possible to employ a reproduction system whose frequency characteristic compensates for that of the record.

It should be pointed out that an attempt to record notes lower than the low cutoff of the above mentioned apparatus would result in recording only those harmonics of the notes which lie above the cut-off. This in no way prevents the listener from hearing the notes, reproduced by means of the harmonics only, as notes with the pitches of the missing

fundamentals although it does somewhat change the quality of the tone.[3] It it were not for this ability of the ear to add the fundamental pitch of a note, of which only the harmonics are being reproduced, most of the older phonographs and loud speakers would have been totally useless for the reproduction of speech and music.

Mechanical Versus Electrical Recording

In attacking the recording part of the problem, two ways at once present themselves; first, the direct use of the power of the sound being recorded to operate the recording instrument; and second, the use of high quality electric apparatus with vacuum tube amplifiers in order to give more freedom to the artists and better control to the process. The amount of power available to operate the recorder directly from the sound in the recording room is so small as to make it extremely difficult to make records under natural conditions of speaking, singing,



Fig. 1a—Picture of an orchestra recording by the acoustic process. This picture was furnished through the courtesy of the Victor Talking Machine Company, Camden, New Jersey

or instrumental playing. As the use of high quality electric apparatus with associated amplifiers has a very distinct advantage over the acoustic method, they have been adopted for the recording part of the process. Fig. 1a shows a picture of a group of artists recording by

[3] Physical Criterion for Determining the Pitch of a Musical Tone, H. Fletcher *Phys. Rev.*, Vol. 23, No. 3, March, 1924.

means of the sound power directly, while Fig. 1b shows a record being made by the same artists with the electric process.

It will be noticed in Fig. 1a that the artists are grouped very closely about the horn. In the case of the weaker instruments such as violins, it has been possible to use only two of standard construction. The rest of the violins are of the type known as the "Stroh" violin which is a device strung in the manner of a violin but so arranged that the bridge



Fig. 1b—Picture of the same orchestra shown in Fig. 1a, but recording by the electric process. This picture was furnished through the courtesy of the Victor Talking Machine Company, Camden, New Jersey

vibrates a diaphragm attached to a horn. The horn is directed toward the recording horn, as shown by the player in the foreground.

With such an arrangement of musicians, it is very difficult to arouse the spontaneous enthusiasm which is necessary for the production of really artistic music. In Fig. 1b the musicians are sitting at ease more nearly in their usual arrangement and all are using the instruments which they would use were they playing at a concert. Furthermore, the microphone is now sufficiently far away from the orchestra to receive the sound in much the manner that the ears of a listener in the audience would receive it. In other words, it picks up the sound after it has been properly blended with the reflections from the walls of the room. It is in this way that the so-called "atmosphere" or "room-tone" has been obtained.

In the old process, it sometimes happened that after the instruments

had been arranged in such a manner that the relative loudness of the various parts had been balanced correctly, it was found that the whole selection was either too loud or too weak. This usually meant a complete rearrangement of the players. With the flexibility introduced by the use of electrical apparatus including amplifiers, the control of loudness is obtained by simple manipulation of the amplifier system and is in no way related to the difficulties of the relative loudness of one instrument to another. The only problem for the studio director in this case is to obtain the proper balance among the various musical instruments and artists. The advantages derived from this added ease of control are also made manifest in that it is much easier and less tiresome for the artists and it is usually possible to make more records in a given time.

Mechanical Versus Electrical Reproducing

Where the question of reproduction is concerned, the same two alternatives mentioned for recording present themselves, namely, direct use of power derived from the record itself versus the use of electromechanical equipment with an amplifier. In this case, however, the situation is a little different as the power which can be drawn directly from the record is more than sufficient for home use. Since any method of reproducing from mechanical records by electrical means involves the use of a mechanical device for transforming from mechanical to electrical power and a second such device for transforming from electrical back to mechanical power, that is, sound, it is necessary to use two mechanical systems, one at each end of an electrical system. Where the power which can be supplied by the record, is sufficient to produce the necessary sound intensity, as in the case of home use, it is in general simpler to design one single mechanical transmission system than it is to add the unnecessary complications of amplifiers, power supply and associated circuits. In cases where music is to be reproduced in large auditoriums, the power which can be drawn from the record may be insufficient and some form of electric reproduction using amplifiers becomes necessary.

Brief Description of Recording System

The system used for recording consists of a condenser transmitter, a high quality vacuum tube amplifier and an electromagnetic recorder. Fig. 2 shows the calibration of the condenser transmitter and the associated amplifiers. The condenser transmitter and amplifiers are so designed that the current delivered to the recorder circuit is essentially proportional to the sound pressure at the transmitter diaphragm. The

electromagnetic recorder, which will be described later, is designed to work with this type of system. With the exception of this electromagnetic recorder, apparatus of this type has already been described in the literature.[4] In addition to this equipment which might be called the

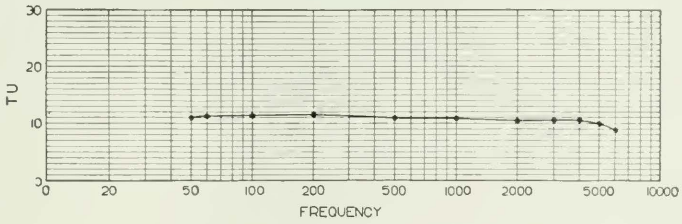


Fig. 2—Calibration of the condenser transmitter and associated amplifiers
This curve shows merely the relative frequency sensitiveness of the system, the zero line having been chosen arbitrarily.

recording amplifier system, there is a volume indicator for measuring the power which is being delivered to the recorder and also an audible monitoring system. The audible monitoring system consists of an amplifier whose input impedance is high compared with the recorder impedance and a suitable loud speaking receiver. The monitoring am-

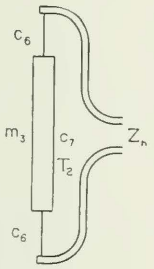


Fig. 3—Schematic mechanical arrangement of diaphragm and air chamber

plifier is bridged directly across the recorder and operates the loud speaking receiver so that the operator may listen to the record as it is being made.

[4] Wente, E. C., "Condenser Transmitter as a Uniformly Sensitive Instrument for Measuring Sound Intensity," *Phys. Rev.*, Vol. 10, 1917.

Crandall, I. B., "Air-Damped Vibrating Systems," *Phys. Rev.*, Vol. 11, 1918.

Wente, E. C., "Electrostatic Transmitter," *Phys. Rev.*, Vol. 19, 1922.

Martin, W. H. and Fletcher H., "High Quality Transmission and Reproduction of Speech and Music," Trans. *A. I. E. E.*, Vol. 43, 1924, p. 384.

Green, I. W. and Maxfield, J. P., "Public Address Systems," Trans. *A. I. E. E.*, Vol. 43, 1923, p. 64.

In the design of the recording and reproducing systems each part of the system has been made as nearly perfect as possible. Errors of one part have not been designed to compensate for inverse errors in another part. Although this method is the more difficult, its flexibility, par-

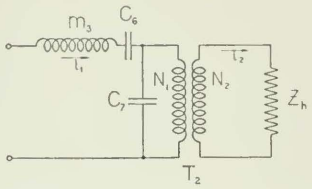


Fig. 4—Electrical equivalent of mechanical system shown in Fig. 3

ticularly as regards the commercial possibilities of future improvements justifies the extra effort.[5] There is, therefore, no distortion in the record whose purpose is to compensate for errors in the reproducing equipment; the only intended distortion in the record being that re-

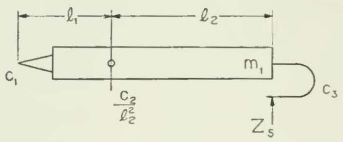


Fig. 5—Schematic mechanical arrangement of needle arm transformer

quired by the inherent limitations mentioned above. See Figs. 2, 14 and 20.

General Basis of Design

An interesting feature of the development of the mechanical and electromechanical portions of the recording and reproducing system is their quantitative design as mechanical analogs of electric circuits. Both the recording and reproducing systems are good examples of the use of this type of analogy.

The economic need for the solution of many of the problems connected with electric wave transmission over long distances coupled with the consequent development of accurate electric measuring apparatus has led to a rather complete theoretical and practical knowledge of electrical wave transmission. The advance has been so great that the knowledge of electric systems has surpassed our previous engineering

[5] Green, I. W. and Maxfield, J. P., "Public Address Systems," Tras. A. I. E. E., Vol. 42, 1923, p. 64.

knowledge of mechanical wave transmission systems. The result is, therefore, that mechanical transmission systems can be designed more successfully if they are viewed as analogs of electric circuits.

While there are mechanical analogs for nearly every form of electrical circuit imaginable, there is one particular class of electrical circuits

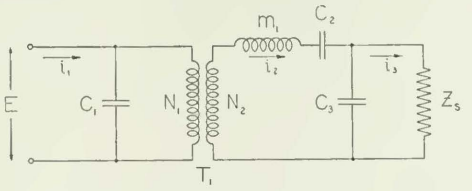


Fig. 6—Electrical equivalent of system shown in Fig. 5 with its termination

whose study has led to ideas of the utmost value in guiding the course of the present development. This class of circuits consists of infinitely repeated similar sections of one or more lumped capacity and inductance elements in series and shunt and are commonly known as filters. The study of filters began with the work of Campbell[6] and a recognition

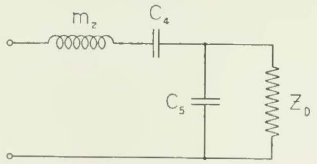


Fig. 7—Electrical equivalent of the spider section

of their importance as frequency selective systems in telephone repeaters, carrier systems, radio, signalling systems, etc., led to their intensive study. In the available literature is to be found a fairly complete statement of their properties and details of their design.[6]

[6] Campbell, G. A., "On Loaded Lines in Telephonic Transmission," *Phil. Mag.*, March 1903.

Campbell, G. A., U. S. Patents 1,227,113; 1,227,114; "Physical Theory of the Electric Wave Filter," *Bell System Technical Journal*, November 1922.

Zobel, O. J.,"Theory and Design of Uniform and Composite Electric Wave Filters," *Bell System Technical Journal*, January 1923.

Peters, L. J., "Theory of Electric Wave Filters Built up of Coupled Circuit Elements," Trans. *A. I. E. E.*, May 1923.

Carson, J. R. and Zobel, O. J., "Transient Oscillations in Electric Wave Filters," *Bell System Technical Journal*, July 1923.

Zobel, O. J., "Transmission Characteristics of Electric Wave Filters," *Bell System Technical Journal*, October 1924.

Johnson, K. S., and Shea, T. E., "Mutual Inductance in Wave Filters with an Introduction on Filter Design," *Bell System Techincal Journal*, January 1925.

Johnson, K. S., "Tansmission Circuits for Telephonic Communication," D. Van Nostrand, 1925.

It will be recalled in the case of the telephone circuit that the introduction of inductance coils at regular intervals in the circuit produced a remarkable change in the transmission characteristic. Over a broad band of frequencies the attenuation was reduced and made fairly uni-

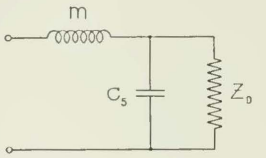


Fig. 8—Electrical equivalent of simple low pass type of network which occurs frequently in this work

form over that range while beyond a critical frequency called the cut-off frequency the attenuation became very high. In the ideal filters with zero dissipation the transmission characteristics are of the same nature but more clear cut. Structures of this type with infinitely repeated sections will have one or more transmission bands of zero attenuation and one or more bands having infinite attenuation. The impedance characteristics of such a structure measured from certain characteristic points will be pure resistance more or less uniform in the transmission bands, and pure reactance in the attenuation bands. These terminations are mid-series; that is, the entering element being one-half of the normal series element; or mid-shunt; that is, the entering element being twice the impedance of the normal shunt element. The corresponding impedances are called the mid-series and mid-shunt characteristic or iterative impedances.

If we retain the first few sections of such a structure and terminate them with a resistance which is equal to the resistance impedance of the infinite line from which they were taken, the characteristics are substantially unchanged. It is understood, of course, that this resistance equals approximately the resistance impedance of the remainder of the infinite line at most of the frequencies in the transmission band in which we are interested.

The presence of small amounts of damping in the various elements also has but slight effect on the general characteristics. These results could in general be readily applied to the various telephone transmission problems because the source and load between which the filter system was inserted generally had or could be made to have a resistance impedance nearly equalling the mid-series or mid-shunt impedance of the filter within the transmission band. The filter and terminating impedances may then be said to be matched. Where adjacent sections

in the filter have impedances similar in character but different in absolute magnitude they may be joined by a suitable transformer.

Many early attempts were made to design mechanical transmission systems having a wide frequency range in which highly damped single or multi-resonant systems were employed. In these attempts both of the obvious methods of increasing the damping were used, namely, that of adding a resistance to the system and that of increasing the value of the compliance and decreasing mass in such proportion as to maintain the same natural frequency. The former of these methods reduces the sensitivity of the system at the point where it is most efficient (See Fig. 9), while the second method increases the response at the points where the system is less sensitive, namely, away from its resonance point. Fig. 9 shows four curves—first, a singly resonant sys-

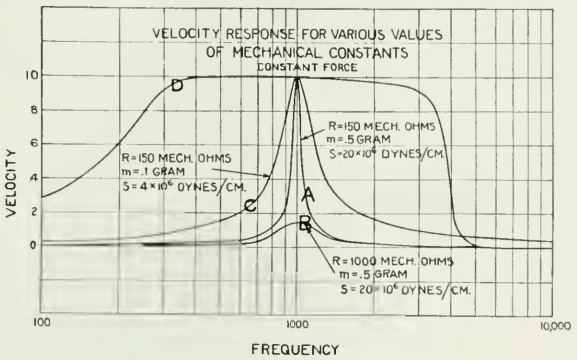


Fig. 9—Velocity response for various values of mechanical constants

tem, Curve *A*; second, the same system with friction added, Curve *B*; third, the same system without the added friction but with an increase in compliance and a decrease in mass such that the natural period remains the same, Curve *C*; and fourth, a band pass type of circuit whose resistance impedance is the same as that of the system shown in Curve *A*. (See Curve *D*.)

The results of filter theory have shown how these resonances should be coordinated so that when a proper resistance termination is used high efficiency and equal sensitivity are obtained over a definite band of frequencies by elimination of response to all frequencies outside the band. With the electrical case of a repeated filter, each section considered by itself resonates at the same frequency but when combined into a short-circuited filter of n sections, there will be n natural frequencies. However, when such a system is terminated with a resist-

ance which equals the nominal characteristic impedance in the transmission band, uniform response in the terminating resistance is obtained over the entire band.

Detailed Analysis of Mechanical and Electrical Analogs [7]

Before going on with a detailed treatment of the electrical analogs of the mechanical structures used in the problem of phonographic reproduction, a list of the corresponding quantities used in the two systems will be given, together with the symbols employed.

Mechanical			Electrical		
Force	= F	(dynes)	Voltage	= E	(volts)
Velocity	= v	(cm./sec.)	Current	= i	(amperes)
Displacement	= s	(cm.)	Charge	= q	(coulombs)
Impedance or mechanical ohms	= z	(dyne sec./cm.)	Impedance	= Z	(ohms)
Resistance	= r	(dyne sec./cm.)	Resistance	= R	(ohms)
Reactance	= x	(dyne sec./cm.)	Reactance	= X	(ohms)
Mass	= m	(gms.)	Inductance	= L	(henries)
Compliance	= c	(cm./dyne)[8]	Capacity	= C	(farads)

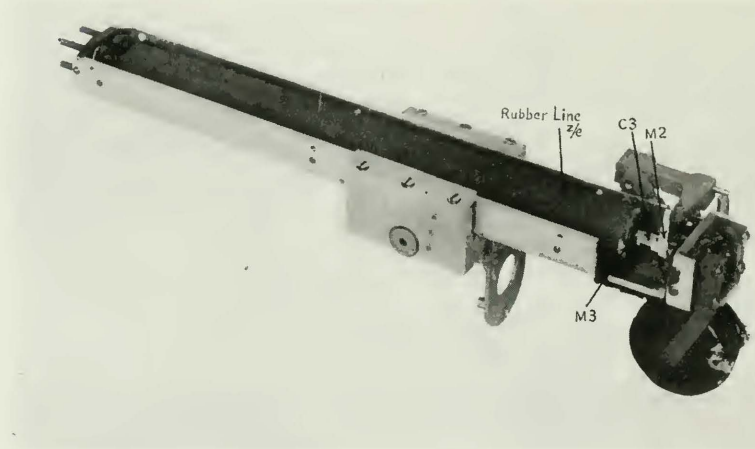


Fig. 10—This figure shows an electromagnetic recorder complete except for the bottom of the case

In addition to the above certain other quantities such as angular displacement, pressure and impedance per unit area, and a few others which have no direct electrical analog will be used. These quantities,

[7] The authors wish to express their appreciation to Mr. E. L. Norton for his courtesy in working out the mathematics of the mechanical and electrical analogs which are shown in this paper.

[8] H. W. Nichols, "Theory of Variable Dynamical Electrical Systems," *Phys. Rev.* Vol. 10, 1917.

however, are either standard in the literature or may always be reduced to those given above.

As illustrations of the general methods employed certain important portions of the reproducer will be considered in detail. Considering first the electrical analog of the air chamber [9] between the diaphragm and horn, we make use of the following list of symbols (see Figs. 3, 4, 15 and 16):

m_3 = Effective mass of diaphragm in grams;
A_1 = Equivalent area of diaphragm in cms^2;
c_6 = Compliance of edge of diaphragm;
c_7 = Compliance of air chamber;
A_2 = Area of throat of horn;
z_h = Impedance of horn—Vector ratio of applied force at the throat of the horn to the resultant linear velocity of the air;
s_1 = Displacement of diaphragm;
v_1 = Velocity of diaphragm;
s_2 = Displacement of air in throat of horn;
v_2 = Velocity of air in throat of horn;
P_0 and V_0 = Initial pressure and volume of air-chamber;
F = Force applied to diaphragm;
p = Small change of pressure in air-chamber.

For a small change p in the pressure within the air-chamber we have:

$$p = \frac{n(A_1 s_1 - A_2 s_2) P_0}{V_0} \tag{1}$$

where $n = 1$ for an isothermal change and 1.4 for an adiabatic change. For the case under consideration $n = 1.4$ very nearly.

If the horn opening is closed, $s_2 = 0$, and we get for the compliance of the air chamber as measured from the diaphragm

$$c_7 = \frac{s_1}{pA_1} = \frac{V_0}{nP_0 A_1^2}.$$

We have the two force equations

$$m_3 \frac{dv_1}{dt} + \frac{s_1}{c_6} + pA_1 = F \tag{2}$$

$$z_h v_2 - pA_2 = 0 \tag{3}$$

[9] The use of the air chamber to increase the loading effect of the horn on the diaphragm has been appreciated for a number of years. It has been used in telephone receivers, phonographs, and loud speaking receivers since their earliest developments. A treatment of the force equations of the air-chamber was given by Hanna & Slepian, "The Function and Design of Horns for Loud Speakers." Trans. *A. I. E. E.*, 1924, p. 393. The equivalent structure, however, was analysed as a compliance and resistance in series instead of in shunt.

or substituting the values of p and c_7

$$m_3\frac{dv_1}{dt}+\frac{s_1}{c_6}+\frac{1}{c_7}\left[s_1-\left(\frac{A_2}{A_1}\right)s_2=F\right] \tag{4}$$

$$z_h v_2+\frac{1}{c_7}\left[\left(\frac{A_2}{A_1}\right)^2 s_2-\left(\frac{A_2}{A_1}\right)s_1\right]=0 \tag{5}$$

If $v_1=j\omega s_1$, etc.

$$z_1 v_1-z_m v_2=F$$
$$z_2 v_2-z_m v_1=0$$

where

$$z_1=j\left(\omega m_3-\frac{1}{\omega c_6}-\frac{1}{\omega c_7}\right),$$

$$z_2=\left[z_h-j\left(\frac{A_2}{A_1}\right)^2\frac{1}{\omega c_7}\right],$$

$$z_m=-j\left(\frac{A_2}{A_1}\right)\frac{1}{\omega c_7}.$$

Considering now the analogous electrical circuit, and assuming the velocity, current, force and voltage to vary sinusoidally, we have the parallel relationship for the steady state conditions:

$$L_3\frac{di_1}{dt}+\frac{q_1}{C_6}+\frac{1}{C_7}\left[q_1-\left(\frac{N_2}{N_1}\right)q_2\right]=E,$$

$$Z_h i_2+\frac{1}{C_7}\left[\left(\frac{N_2}{N_1}\right)^2 q_2-\left(\frac{N_2}{N_1}\right)q_1=0.\right.$$

where $\frac{N_2}{N_1}$ = turns ratio of ideal transformer (Fig. 4).

If $i_1=\mathrm{j}\omega q_1$, etc.

$$Z_1 i_1-Z_m i_2=E$$
$$Z_2 i_2-Z_m i_1=0$$

where

$$Z_1=j\left(\omega L_3-\frac{1}{\omega C_6}-\frac{1}{\omega C_7}\right),$$

$$Z_2=\left[Z_h-\mathrm{j}\left(\frac{N_2}{N_1}\right)^2\frac{1}{\omega C_7}\right],$$

$$Z_m=-j\left(\frac{N_2}{N_1}\right)\frac{1}{\omega C_7}.$$

The last five equations in each case give the complete solution of the network. By analogy between the two sets of equations, therefore, the air-chamber corresponding in the electrical case to a shunt capacity, c_7 is spoken of as a shunt compliance, $c_7 = \frac{V_0}{nP_0A_1^2}$, together with a transformer inserted before the horn, which has an equivalent turns ratio equal to the ratio of the areas of the diaphragm and horn openings.

Taking up now the somewhat different illustration of the needle arm, the following symbols are needed (Figs. 5, 6, 15, 16):

l_1 = Distance from pivot point to end of needle;
l_2 = Distance from pivot point to center of "spider" (Fig. 15);
I = Moment of inertia of needle arm;
m_1 = Apparent or equivalent mass of arm as measured from the center of the spider

$$= \frac{I}{l_2^2};$$

c_1 = Compliance of needle point;
c_2 = Compliance of bearing to turning of the needle arm, as measured from end of arm at the spider;
c_3 = Compliance of end of needle arm attached to spider;
s_1 = Displacement of tip of needle;
s_2 = Displacement of end of arm at the spider;
s_3 = Displacement of spider;
z_s = Mechanical impedance of spider and remainder of structure = Vector ratio of applied force to resultant velocity;
θ = Angular displacement of needle arm;
F = Applied force at needle point.

We have the three force equations:

$$\frac{s_1 - l_1\theta}{c_1} = F \tag{6}$$

$$I\frac{d^2\theta}{dt^2} + \frac{(l_1\theta - s_1)l_1}{c_1} + \frac{\theta l_2^2}{c_2} + \frac{(l_2\theta - s_3)l_2}{c_3} = 0 \tag{7}$$

$$\frac{s_3 - l_2\theta}{c_3} + z_s\frac{ds_3}{dt} = 0 \tag{8}$$

Replacing θ by $\frac{s_2}{l_2}$ and I by $m_1 l_2^2$ gives:

$$\frac{s_1 - \frac{l_1}{l_2} s_2}{c_1} = F \tag{9}$$

$$m_1 \frac{d^2 s_2}{dt^2} + s_2 \left[\left(\frac{l_1}{l_2} \right)^2 \frac{1}{c_1} + \frac{1}{c_2} + \frac{1}{c_3} \right] - \frac{l_1}{l_2} \frac{s_1}{c_1} - \frac{s_3}{c_3} = 0 \tag{10}$$

$$\frac{s_3 - s_2}{c_3} + z_s \frac{ds_3}{dt} = 0 \tag{11}$$

Considering now the parallel mechanical electrical circuits, and assuming as before sine functions for v, i, F, and E, we have:

Mechanical Case, substituting $v_1 = \mathrm{j}\ \omega\ s_1$, etc., in the last equations:

$$-j \frac{v_1}{\omega c_1} + j \frac{l_1}{l_2} \frac{v_2}{\omega c_1} = F,$$

$$j v_2 \left[\omega m_1 - \left(\frac{l_1}{l_2} \right)^2 \frac{1}{\omega c_1} - \frac{1}{\omega c_2} - \frac{1}{\omega c_3} \right]$$

$$+ j \frac{l_1}{l_2} \frac{v_1}{\omega c_1} + j \frac{v_3}{\omega c_3} = 0,$$

$$j \frac{v_2}{\omega c_3} + v_3 \left(z_s - j \frac{1}{\omega c_3} \right) = 0.$$

Electrical Case, with ideal transformer of turns ratio $\frac{N_2}{N_1}$:

$$-j \frac{i_1}{\omega C_1} + j \left(\frac{N_2}{N_1} \right) \frac{i_2}{\omega C_1} = E,$$

$$j i_2 \left[\omega L_1 - \left(\frac{N_2}{N_1} \right)^2 \frac{1}{\omega C_1} - \frac{1}{\omega C_2} - \frac{1}{\omega C_3} \right] + j \left(\frac{N_2}{N_1} \right) \frac{i_1}{\omega C_1} + j \frac{i_3}{\omega C_3} = 0,$$

$$j \frac{i_2}{\omega C_3} + i_3 \left(Z_s - j \frac{1}{\omega C_3} \right) = 0.$$

The analogy between the two sets of equations is quite obvious. It will be noticed that the effect of the lever arm is to introduce an equivalent transformer of a turns ratio which is the reciprocal of the corresponding lengths of the arms.

The general method of deducing the equivalent electric circuits should be clear from the above illustrations of the air-chamber and

of the needle arm. For example, in the spider section, Fig. 15, the mass is driven directly by the force from the needle-arm compliance, there being a small series compliance in the connection owing to bending of connecting rod. The diaphragm is connected through the

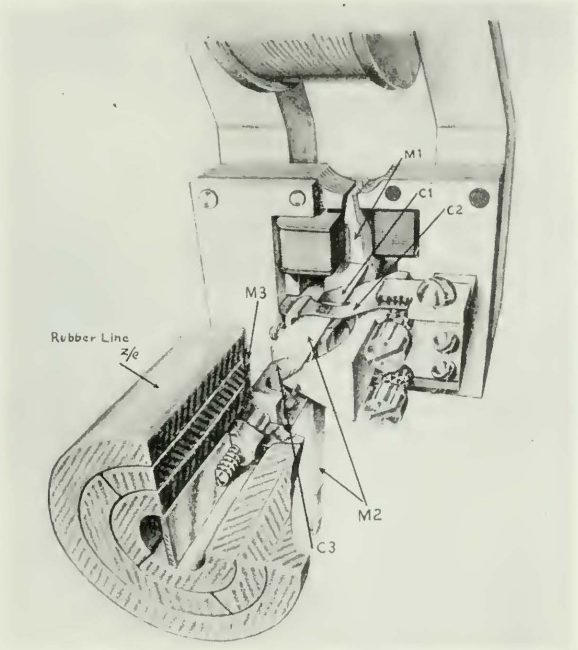


Fig. 11—Detailed drawing of the mechanical filter of an electromagnetic recorder. (Lettering same as in Fig. 12)

compliance of the prongs of the spider. The equivalent circuits are shown in Figs. 7 and 16.

The equations of this network may be obtained from the equations for the needle arm by placing c_1 equal to zero, taking a unity ratio transformer, and substituting m_2 for m_1, c_4 for c_2, c_5 for c_3 and z_d for z_s.

Another type of network which occurs frequently in the building of mechanical vibrating systems is represented diagrammatically in Fig. 8. This is clearly a particular case of Fig. 7 with c_4 made infinite.

By combining Fig. 6 representing the needle arm, Fig. 7, representing the spider section and Fig. 4 representing the diaphragm, air-chamber and horn, the complete reproducer may be built up. The resultant network is shown in Fig. 16. Since methods are available in the theory of electric wave filters to determine the proper

values of the elements of the complete network for a free transfer of energy throughout an assigned frequency band, the analogous mechanical elements may be determined in the same manner.

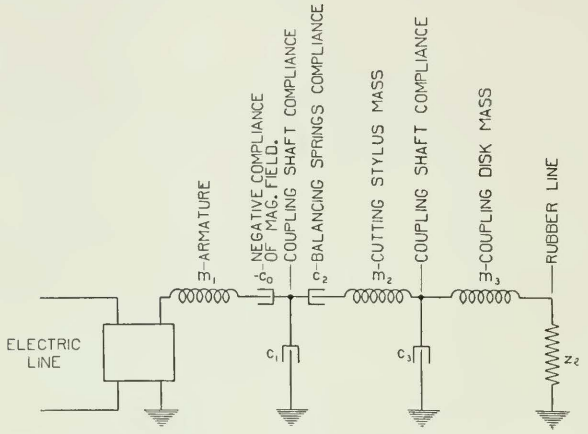


Fig. 12—Equivalent electric circuit of the electromagnetic recorder

General Design of Mechanical Systems

In designing mechanical systems of the band pass type, the problem is three fold—first, that of arranging the masses and compliances such that they form repeated filter sections; second, determining the magnitude of these quantities so that with or without transformers the separate sections all have the same cut-off frequenceis[10] and characteristic impedances; third, to provide the proper resistance termination. Where the transmitted mechanical power has not been radiated as sound this third part has been one of the most difficult to fulfill.

In designing these systems, practical difficulties arose—first, the difficulty of insuring that the parts vibrated in the desired degrees of freedom only, and second, the difficulty of determining the magnitudes of the various effective masses, compliances and resistances. Before the work to be described could be carried out practically it became necessary to develop a method of measuring mechanical impedances[11].

[10] It is of course permissible to have a section having a higher cut-off than the others provided its characteristic impedance is the same as that of the others over the transmission band of those having the lower cut-off.

[11] Kennelly, A. E. and Affel, H. A., "The Mechanics of Telephone Receiver Diaphragms, as Derived from their Motional Impedance Circles," *Proc. A. A. A. S.*, Vol. 51, No. 8, November, 1915.

Kennelly, A. E. and Pierce, G. W., "The Impedance of Telephone Receivers as Affected by the Motion of their Diaphragms," *Proc. A. A. A. S.*, Vol. 48, No. 6, September, 1912.

Such a method has been developed which at the present time covers a range of frequencies from somewhere below 50 to about 4,500 pps. Work is still being continued to extend the method to the higher frequencies. This method of measurement has been very useful not

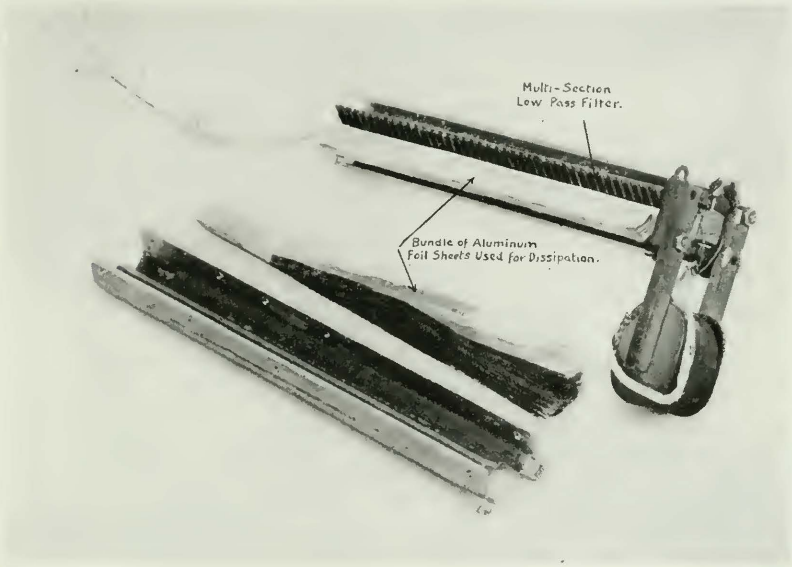


Fig. 13—Electromagnetic recorder using lumped loaded termination
The method of furnishing dissipation to the lumped loaded line is shown

only in determining the magnitudes of the impedances in the degrees of freedom in which it is desired that they shall operate, but in determining the impedances to motion of the various parts in directions in which they should not be permitted to vibrate. In connection with the measurement of the magnitudes of the parts in the desired degrees of freedom this method enables us to determine the constants of the mechanical networks under their conditions of operation. Experience so far has indicated that when all the degrees of freedom have been taken into account and when the dynamic axes of vibration have been properly chosen, the static and dynamic constants of the parts are the same, and it is then possible to check the parts by simple static measurements. In the early attempts to build these systems very large discrepancies between the static and dynamic characteristics were found.

The Recorder

One of the early practical phonographic applications of electric filter design to mechanical problems was the development of an electromagnetic recorder. The instrument as finally constructed is essentially a properly terminated three-section mechanical filter in which the recording stylus and its holder constitute the series mass in the second section. Since a filter of this type appears at its input end as approximately a pure resistance within the transmission band, the current in the series inductances, that is, in the mechanical case, the velocity of the series masses is proportional to the driving force.

Figs. 10, 11 and 12 show respectively, a complete recorder, a drawing of the mechanical filter of such a recorder and a diagram of the equivalent electric circuit. The armature acts as the series mass m_1 in the first section; the magnetic field as the series negative compliance, $-c_0$; the shaft between the armature and the stylus holder as the shunt compliance c_1; the balancing springs as the series compliance c_2; the stylus holder and the stylus as the series mass m_2; the shaft between the stylus holder and the disk, coupling the system to the terminating resistance, as the compliance c_3; the coupling disk as the series mass m_3 and the terminating line as approximately a mechanical resistance.

All of these equivalents are seen from the simple analog previously outlined with the exception of the terminating resistance and the negative compliance, $-c_0$. The terminating resistance was originally made up of a series of filter sections of lumped series masses and shunt compliances with a small amount of damping added to the motion of each of the series masses. Fig. 13 shows one of the early recorders equipped with this type of resistance termination. The reason for using such a complicated termination lies in the fact that most of the known mechanical resistances have values which are functions of frequency or of amplitude or both. Also in most cases, the mechanical resistance is accompanied by either a mass or compliance reactance. By using a multi-section filter which is sufficiently long so that a wave entering it will be essentially absorbed before it has reached the far end, been reflected and returned to the entering end, it has been possible to use imperfect types of damping for this line and still obtain over the desired band, an essentially pure resistance at the input end.

More recently a continuous line has been developed which is much easier of practical attainment than the complicated lump-loaded filter. The recorder shown in Fig. 10 is so equipped.

Fig. 14 shows calibration curves of three types of recorders. The

bottom curve shows an early type of highly damped singly resonant system. The middle curve is a calibration of a low pass mechanical filter type using lumped loading in the resistance line. The upper curve shows the calibration of the recorder shown in Fig. 10.

The compliance $-c_0$ is a mechanical quantity for which there is no simple electric analog. In a balanced armature type of structure such as that shown in Fig. 11, the action of the field on the armature, when it is at its center point, is balanced. If, however, the armature be de-

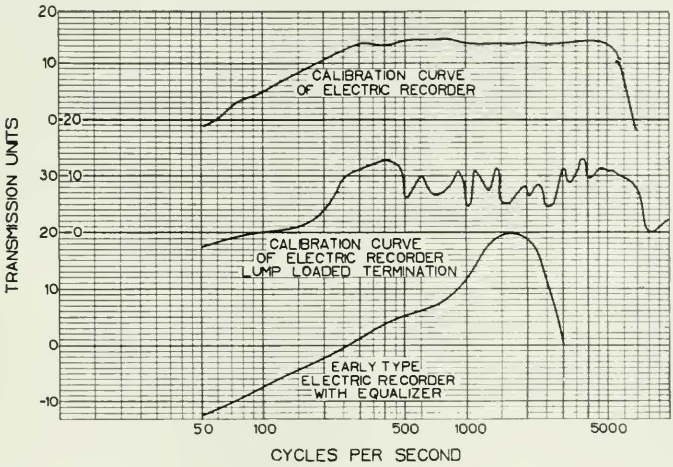


Fig. 14—Calibration curve of three types of electromagnetic recorders

flected, a small distance from this equilibrium, there is exerted by the magnetic field a torque tending to pull the armature further away from its center position. The value of this torque for small amplitudes is proportional to the angular displacement. It is therefore seen that this quantity is of the nature of a compliance but that the back force is in a reverse direction to that required for a positive compliance.

Design of the Reproducing Apparatus

As the analogy between the mechanical and electrical filter is more perfectly shown in the case of the reproducing equipment, its detailed quantitative description will now be given. Figs. 15 and 16 show respectively a diagram of the reproducing system and its equivalent electric circuit. From these diagrams it is evident which units in the mechanical system correspond to the various electrical parts. As the series compliances c_2, c_4 and c_6 have been made so large that

the low frequency cut-off caused by them lies well below the low frequency cut-off of the horn, an inappreciable error is introduced in

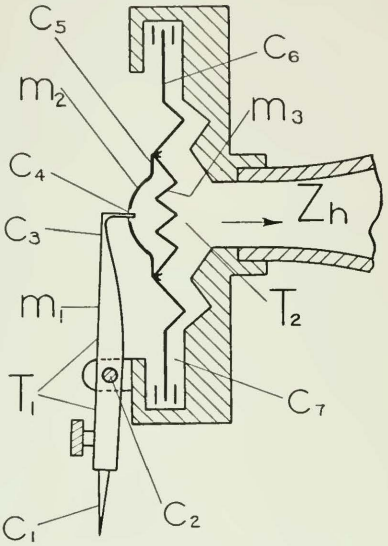


Fig. 15—Diagrammatic sketch of the mechanical system of the phonograph

using for design purposes formulas of low pass filters[12]. The two formulas which will be used are as follows:

$$f_c = \frac{1}{\pi}\sqrt{\frac{1}{mc}} \tag{12}$$

Where

f_c = cut-off frequency of a lumped transmission system in cycles per second

c = shunt compliance per section in centimeters per dynes

m = series mass per section in grams

$$z_o = \sqrt{\frac{m}{c}} \tag{13}$$

where z_0[13] is the value of characteristic impedance over the greater part of the band range.

[12] Campbell, G. A., "On loaded lines in Telephonic Transmission," *Phil. Mag.*, March, 1903.

[13] z_0 may be called nominal mid-shunt or mid-series impedance. Their actual values in the transmission band being at any frequency f.,

$$\text{mid-series} = z_0\sqrt{1-\left(\frac{f}{f_c}\right)^2} \qquad \text{mid-shunt} = \frac{z^0}{\sqrt{1-\left(\frac{f}{f_c}\right)^2}}$$

Equations (12) and (13) which form the basis of the design work contain four variables, f_c, c, m and z_0. It is, therefore, necessary to determine two of them by the physical requirements of the problem after which the other two are determined. The upper cut-off frequency f_c was arbitrarily chosen at 5000 pps. as a compromise between the highest frequency occurring on the record and the increase in surface noise

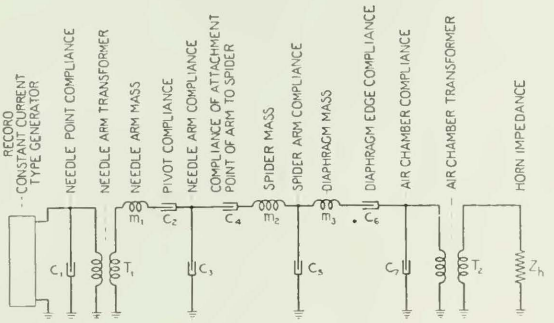


Fig. 16—Electric equivalent of the system shown in Fig. 15

as the cut-off is raised. The choice of the other arbitrarily set variable came after considerable preliminary experimenting and was fixed by the difficulty of obtaining a diaphragm which is light enough and has a large enough area. Hence the effective mass of the diaphragm m_3, (Figs. 15–16) was fixed at 0.186 grams which value can be obtained by careful design. The effective area can be made as large as 13 square centimeters. For convenience let the arbitrary value chosen for $f_c = \bar{f}_c$ and the value of $m = \bar{m}_3$.

Solving Equations (12) and (13) for c and z_0, we get

$$c = \frac{1}{\pi^2 \bar{f}_c^2 \bar{m}_3}, \tag{14}$$

$$z_0 = \pi \bar{f}_c \bar{m}_3; \tag{15}$$

also

$$z_0 = \frac{1}{\pi c \bar{f}_c} \tag{16}$$

In order to obtain the low value of mass mentioned, with a large enough area, it was necessary to make the diaphragm of a very stiff light material. An aluminum alloy sheet 0.0017 in. thick was chosen and concentrically corrugated as shown in Figs. 17 and 18. These corrugations are spaced sufficiently close so that the natural periods of the

flat surfaces are all above $\bar{f}_c$. To insure that this central stiffened portion should vibrate with approximate plunger action, which is more efficient than diaphragm action, it is driven at six points near its periphery.

Reference to Figs. 15 and 16 and Equation (14) shows that the compliance of the air chamber c_7, of the spider legs c_5 and shunt tip of the needle arm c_3 are determined. Also the mass of the spider m_2 and the effective mass of the needle arm m_1, as viewed at the point where it is attached to the spider, are determined.

The impedance looking into the system from the record is determined by the rate at which it is necessary to radiate energy in order that the reproduction may be loud enough. The power taken from the record is approximately $v^2\, z_0$ since z_0 is a resistance over most of the band. Experiment has shown this value of z_0 to be approximately 4500 mechanical ohms.

But substituting in Equation (13) the value of $\overline{m}_3$, and from Equation (14) the value of c_5, we find that the impedance is only 2920 mechanical ohms. It is, therefore, necessary to use a transformer whose impedance ratio is $\frac{4500}{2920}$. From this and a knowledge of filter structures the needle-point compliance can be determined. The value obtained is easily realized with commercial types of needle.

It will be noted that the record is shown in Fig. 16 as a constant current generator, *i. e.*, a generator whose impedance appears high as viewed from the needle point. That this is necessary is obvious when it is remembered that, if the impedance looking back into the record were to equal the impedance of the filter system, the walls of the record would have to yield an amount comparable with one-half the amplitude of the lateral cut. This would cause a breakdown of the record material with consequent damage.

The design of the system is, therefore, complete except for the resistance termination which is supplied by the horn for all frequencies above its low frequency cut-off. The characteristics of the horn will be dealt with later. The resistance within the band looking in at the small end of the horn is $G\, A_2$ where G equals the mechanical ohms per square centimeter of an infinite cylindrical tube of the same area, and A_2 equals the area in square centimeters of the small end of the horn.

Let A_1 = the effective plunger area of the diaphragm (as previously mentioned this is 13 sq. cm.). The impedance looking back at the diaphragm is

$$z_0 = \pi \bar{f}_c\, \overline{m}_3 = 2920 \text{ mechanical ohms}$$

from Equation (15), and the impedance looking at a horn whose small end area equals A_2 is

$$z_h = r_0 = A_2 G \tag{17}$$

Substituting

$A_2 = 13$ sq. cm.
$G = 41$ ohms per cm.^{2}

we get

$z_h = r_0 = 533$ mechanical ohms

This is entirely insufficient so that the air-chamber transformer becomes necessary.

To calculate the necessary ratio of areas on the two sides of the air-chamber transformer, the following formula is needed. The formula assumes the chamber to be relatively small compared with all wave lengths of the sound to be transmitted, that is, the pressure changes throughout the chamber are substantially in phase.

$$\frac{z_0}{z_h} = \left(\frac{v_2}{v_1}\right)^2 = \left(\frac{F_1}{F_2}\right)^2 = \left(\frac{A_1}{A_2}\right)^2 \tag{18}$$

where

z_0 = the impedance of the primary side of the transformer in mechanical ohms;
z_h = the impedance on the secondary side of the transformer in mechanical ohms, *i.e.*, the horn impedance;
v_1 = mechanical current, *i.e.*, velocity on the primary side of the transformer in centimeters per second;
v_2 = mechanical current on the secondary side of the transformer in centimeters per second;
F_1 = alternating force on primary side of air-chamber transformer in dynes;
F_2 = alternating force on secondary side of air-chamber transformer in dynes;
A_1 = effective area working into the primary side of the air-chamber in centimeters squared;
A_2 = effective area working into the secondary side of the air-chamber in centimeters squared.

The characteristic impedance of the line on the diaphragm or primary side of the air-chamber as shown by equation (15) is

$$z_0 = \pi \bar{f}_c \bar{m}_3. \tag{19}$$

From Equation (17) the characteristic impedance on the horn or secondary side is

$$z_h = GA_2. \tag{20}$$

Therefore,

$$\left(\frac{A_2}{A_1}\right)^2 = \frac{z_h}{z_0} = \frac{GA_2}{\pi f_c m_3} \tag{21}$$

and solving this for A_2, we get

$$A_2 = \frac{GA_1^2}{\pi f_c m_3} \tag{22}$$

The equivalence of the air-chamber to a transformer shunted by a compliance is shown earlier in the paper.

In applying the foregoing method of design to a practical structure, a number of design problems had to be solved. The construction of the diaphragm and the method by which it is actuated have been already described, except for the tangential corrugations constituting the series compliance. The use of these corrugations results in the

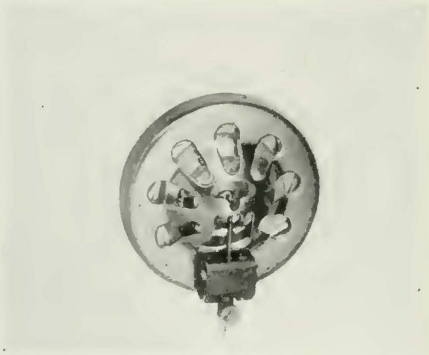


Fig. 17—Photograph of mechanical reproducing system without the horn

value of the series compliance being practically independent of the nature of the clamping, and has eliminated a tendency to "rattle" introduced by unevenness in the clamping surfaces.

Another feature in connection with the sound box is the needle-arm bearing shown in Figs. 17 and 18. Ordinary knife-edge bearings are not sufficiently rigid as fulcrums and the rotational reactance as well as the rotational resistance is undesirably large. A construction which has been found to meet the necessary requirements is the ball bearing type with the steel balls held in position by magnetic pull. By making the ball-containing case of soft steel and magnetizing the shaft, it has been possible to manufacture this bearing reliably and cheaply.

The horn which has been used as a terminating resistance to the mechanical filter structure is a logarithmic one. The general properties of logarithmic horns have been understood for some time.[14]

There are two fundamental constants of such a horn—the first is the area of the large end and the second the rate of taper. The area of the

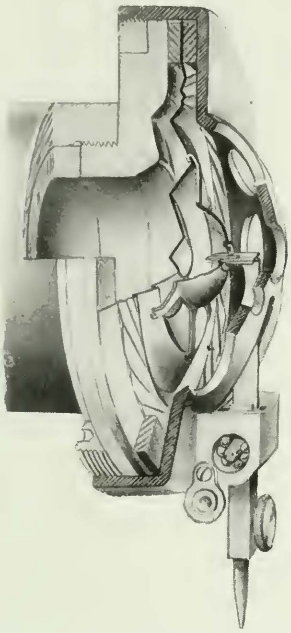


Fig. 18—Sectional drawing showing construction of the system shown in Fig. 17

mouth determines the lowest frequency which is radiated satisfactorily. The energy of the frequencies below this is largely reflected if it is permitted to reach the mouth.

From the equations given by Webster, [14] it can be shown that all logarithmic horns have a low frequency cut-off which is determined by the rate of taper. If the rate of taper is so proportioned that its resulting cut-off prevents the lower frequencies from reaching the horn mouth, the horn will then radiate all frequencies reaching its mouth and very little reflection will result.[15] It is, therefore, possible to build a horn having no marked fundamental resonance.

[14] Webster, A. G., "Acoustical Impedance and Theory of Horns and Phonograph," *Proc.* Nat. Acad. of Sci., 1919.

[15] The authors wish to express their appreciation in this connection of the work of Mr. P. B. Flanders who carried out the mathematical investigation of these relationships and to Mr. A. L. Thuras who checked experimentally the mathematical theory.

Since the characteristics of the horn are determined by the area of its mouth and by its rate of taper the length of the horn is determined by the area of the small end. This area is determined in turn by the mechanical impedance and effective area of the system which it is terminating, as shown in Equation (22). It is seen, therefore, that the length of the horn should not be considered as a fundamental constant. A paper describing the design of horns based on these principles is being prepared.

An interesting feature of the horn which has been built commercially is its method of folding. The sketch in Fig. 19 shows a shadow picture

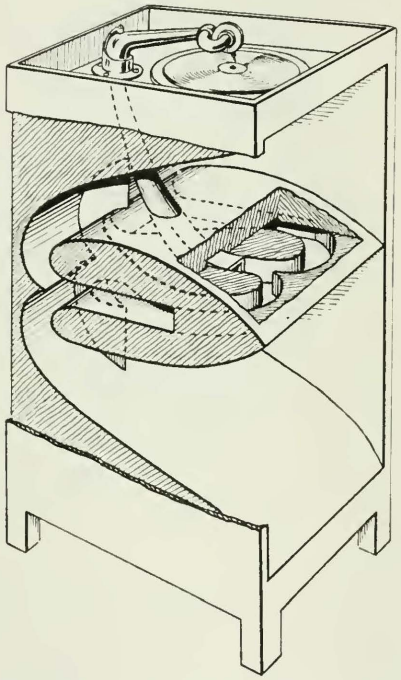


Fig. 19—Sectional view of the folded horn showing the air passage

of the horn. It will be noticed that the sound passage is folded only in its thin direction, which permits the radius of the turns to be small and thereby makes the folding compact.

Fig. 20 shows the frequency characteristic of a phonograph designed as shown above with a logarithmic horn whose rate of taper and area of mouth opening place the low cut-off at about 115 cycles. It also shows the characteristics of one of the best of the old style phonographs. Curve *A* represents the new machine, while Curve *B* repre-

sents the old style standard machine. Since the vertical scale used in this graph is logarithmic the full difference between the two instru-

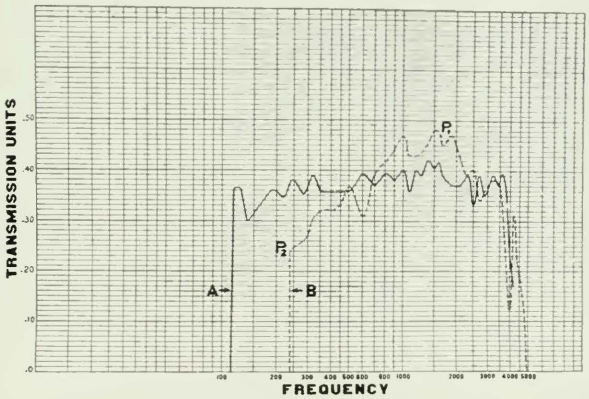


Fig. 20—Response frequency characteristic of two phonographs. Curve *A* shows the characteristic of the band pass filter type described. Curve *B* shows the characteristic of one of the best commercial machines previously on the market



Fig. 21—Bridge for measuring mechanical impedance, being used for determining the impedance of a phonograph horn

ments does not appear at first glance. Some idea as to the magnitude of this difference can be obtained, however, by noting that the points P_1 on the curve of the old machine stands at a power level about 250 times as great as P_2.

Abstracts of Recent Technical Papers from Bell System Sources

Complex Magnetization.[1] EUGENE PETERSON. Magnetization of silicon steel by two sinusoidal fields of differing frequencies. The energy loss W per cycle and the flux density B associated with each of the two frequencies were determined when the two sinusoidal magnetizing forces were simultaneously impressed on a toroidal silicon steel core built up of one-mil laminations. A null method was used which permitted suppression of the modulated currents and constancy of the impressed currents during manipulation for balance. The frequencies used were 400, 821 and 1582. Six sets of measurements were taken with fixed magnetizing forces ranging from 0.5 to 10 gilberts/cm and superposed forces up to 15 gilberts/cm. The results show that the effect of superposition depends upon the relative amplitudes and upon the frequency ratio R of the superposed frequency to the other. At low fixed fields W and B go through maxima as the superposed field is increased, the maximum value increasing with R. The maximum is less pronounced or absent for the higher fixed fields. In general B is smaller with a low than with a high value of R other things being equal. The effect on W is not as sharply defined; in general the effect of superposition is more pronounced the higher the superposed frequency. The amplitude effect and frequency ratio effect are shown to be in general agreement with conclusions drawn from mathematical treatment of somewhat simplified cases and it is concluded that the effects are not inconsistent with purely hysteretic phenomena.

Some Photographic Problems Encountered in the Transmission of Pictures by Electricity.[1] HERBERT E. IVES. This paper considers some of the problems of photographic tone reproduction, which arise upon the introduction of an electrical transmission system between a picture placed on sending apparatus in one place and the copy of the picture made by receiving apparatus in another place. Some of these problems arise because of limitations introduced by the use of the electrical transmission line; others arise because of opportunities for the control of picture quality which are not afforded by ordinary photographic methods. As an illustration of one of these limitations

[1] *Physical Review*, Vol. 27, pp. 318–328, March, 1926.

[2] *Journal of the Optical Society of America and Review of Scientific Instruments*, March, 1926, pp. 173–194.

may be mentioned the fact that the original picture, for instance a photographic negative, is not seen by the operator at the receiving end. He cannot, therefore, by using his photographic knowledge and experience, choose printing media and decide upon conditions of exposure and development. As an illustration of the opportunities introduced by an electrical picture transmission apparatus may be noted the possibility of so poling the electrical elements that the received picture may be either a positive or negative, irrespective of the nature of the original at the sending end.

While in other picture transmission systems other problems arise peculiar to these systems, it is believed that although the questions considered are those presented in commercial operation in the Bell System, they are, to a certain extent, common to all electrical picture transmission apparatus.

A Radio Field-Strength Measuring System for Frequencies up to Forty Megacycles.[3] H. T. FRIIS and E. BRUCE. In previous types of radio field strength measurement apparatus it is very difficult to reproduce accurately the small comparison voltages at very high frequencies, due to reactive effects in the attenuating networks. The "tube voltmeter" is practically the only reliable instrument available at high frequency measurement work. New measurement sets for very high frequency signals have, therefore, been developed. The apparatus is a double detection receiving set which is equipped with a calibrated intermediate frequency attenuator and a local signal comparison oscillator. The local signal is measured by means of the intermediate frequency detector which is calibrated as a tube voltmeter and all required attenuations are made at the relatively low and fixed intermediate frequency.

A New Mechanical Test for Rubber Insulation.[4] C. L. HIPPENSTEEL. This paper discusses the development of a rapid routine test which will numerically express the ability of the rubber insulation to resist cutting by the conductor at the points of support and to resist cracking at points of extreme flexure. Up to the present time no one test of that nature has been described.

[3] Presented at a meeting of the Institute of Radio Engineers, May 5, 1926.

[4] *Industrial and Engineering Chemistry*, April, 1926.

Contributors to this Issue

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 analyzing speech.

C. J. Beck, Western Electric Company Engineering Department, 1917; U. S. Army Air Service, 1918-19.—Since returning to the Laboratories, Mr. Beck has been associated with Mr. Sacia in speech analyses.

J. J. Gilbert, A.B., University of Pennsylvania, 1909; Harvard, 1910-11; Chicago, 1911-12; E.E., Armour Institute, 1915; instructor of electrical engineering, Armour, 1912-17; Captain Signal Corps, 1917-19; Engineering Department, Western Electric Company, 1919; Bell Telephone Laboratories, Inc., 1925—. Mr. Gilbert has worked primarily on submarine cable problems.

R. B. Shanck, B.E.E., Ohio State University, 1915. Railroad Telegraph Service, 1909-10; Plant Department, American Telephone and Telegraph Company, 1910-11 and summers 1912-14; Engineering Department, 1915-19; Department of Development and Research, 1919—. Since 1915 Mr. Shanck has been engaged in transmission development work, on telegraph systems and compositing arrangements.

F. B. Llewellyn, M.E., Stevens Institute of Technology, 1922; graduate student, Columbia University, 1924-5; Research Department, Western Electric Company, 1922-5; Bell Telephone Laboratories, 1925—.

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.

Joseph P. Maxfield, S.B., Massachusetts Institute of Technology, 1910; instructor in electro-chemistry, Massachusetts Institute of

Technology, 1910-12; instructor in physics, Massachusetts Institute of Technology, 1912-14; industrial research work, 1910-14; Western Electric Company, 1914-24; Bell Telephone Laboratories, 1925—. Mr. Maxfield has engaged in fundamental studies of microphonic contacts and electro-acoustic devices. Recently these activities have been directed toward the development of public address systems and methods of recording and reproducing music and speech particularly as regards their application to the phonograph and audible motion pictures.

Henry C. Harrison, A.B., Colorado College, 1910; S.B., Massachusetts Institute of Technology, 1913; instructor in electrical engineering, Massachusetts Institute of Technology, 1913-14; Western Electric Company, 1914-24; Bell Telephone Laboratories, 1925—. Mr. Harrison has made fundamental studies of receivers and carbon button microphones. More recently, his work has been principally concerned with the design and development of various mechanical apparatus to embody the principles of electric transmission theory.

The Bell System Technical Journal

October, 1926

Radio Signaling System for the New York Police Department

By S. E. ANDERSON

Synopsis: By means of the radio signaling system described it is possible, through the addition of a comparatively simple attachment to a standard radio telephone transmitter, to modulate the carrier with an audio frequency tone in such a manner as to provide for calling individually, simultaneously, or in a number of designated groups, any one of several hundred radio receiving stations. At the radio receiving stations apparatus is provided which may be operated from commercial sources of power supply and by means of which a visible or audible signal is given to the operator that a message is about to be broadcast, to which he should listen. Signals are also provided which, in case of improper operation, immediately inform the operator of the points at which attention is required.

Introduction

FOR some time the New York Police Department has been employing the municipal broadcasting station WNYC to broadcast descriptions of missing persons and other features of police work in which it is desired to enlist the cooperation of the public. The success of this program has been such that the Police Department wished to equip the precinct houses and police booths located in various parts of the city with receiving sets with which they could listen in on communications from the headquarters station WNYC. The fundamental requirement was signaling apparatus incorporated in the receiving sets which would attract the attention of the attendant at the proper times. The system which was finally developed by the engineers of the Bell Telephone Laboratories, Inc., in cooperation with the New York Police Department is an excellent illustration of the adaptability of wire practices to the radio field. The underlying principles employed and much of the apparatus used had previously found extensive application in the Bell System and elsewhere.

The basis of this system is the Western Electric telephone train dispatching system which is in rather general use on railroads throughout the world for the purpose of providing selective signaling on their train dispatching telephone systems. For every division, these systems consist ordinarily of a single line to which are connected a number of stations capable of being called by the dispatcher in-

dividually, simultaneously or in groups.[1] This signaling system has also been adapted to radio transmission.[2] Its use permits broadcasting from a central radio transmitting station to police organization districts, patrol boats and automobiles without requiring the constant attention of operators at the receiving stations.

Requirements of the System

Before describing the system which was finally worked out to meet the requirements of the New York Police Department, it seems best to state the nature of these requirements. For a system employed to handle communications ranging all the way from routine messages between police headquarters and its different outlying police stations and patrols to general alarms for insuring the capture of escaping criminals, absolute reliability and flexibility are of the utmost importance. The central station must be able to call the receiving stations individually, collectively, or in a number of designated group combinations corresponding to the police organization districts. To accomplish this result effectively, means must be provided whereby the desired signal may be sent automatically by a simple manual operation. The apparatus for this purpose must be in the form of an attachment which may be used with a standard radio telephone transmitter without extensive modifications.

As the receiver will be in operation continuously, the difficult and expensive maintenance of batteries must be avoided by energizing the vacuum tubes from the commercial power supply system. The tuning arrangements of the receiver must be of the simplest possible character and must be capable of being locked to insure that the receiver remains tuned to the transmitting frequency. The selectivity and sensitivity must be sufficient to insure reliable operation under all conditions. The receiver must provide means for listening to all material broadcast by the central broadcasting station but the signaling system should respond only to signals from the transmitter signaling attachment, irrespective of broadcast speech, music and telegraph signals which may involve the same frequencies. Visible indications should be provided to show when the receiver is in operating condition. The receiver should respond to a call from the central station by operating another visible indicator, in addition to a bell or other audible signal, if desired.

[1] "Modern Methods in Train Dispatching," by J. C. Latham, *Electrical Communication*, Vol. III, No. 1, July, 1924.

[2] "Radio Telephone Signaling Low-Frequency System," by C. S. Demarest, M. L. Almquist and L. M. Clement, *Journal of the A. I. E. E.*, Vol. XLIII, No. 3 March, 1924.

Description of Apparatus

Transmitter Attachment

A photograph of the transmitter attachment is shown in Fig. 1, and a schematic circuit is given in Fig. 2. The apparatus consists of a vacuum tube oscillator and a number of calling keys. These

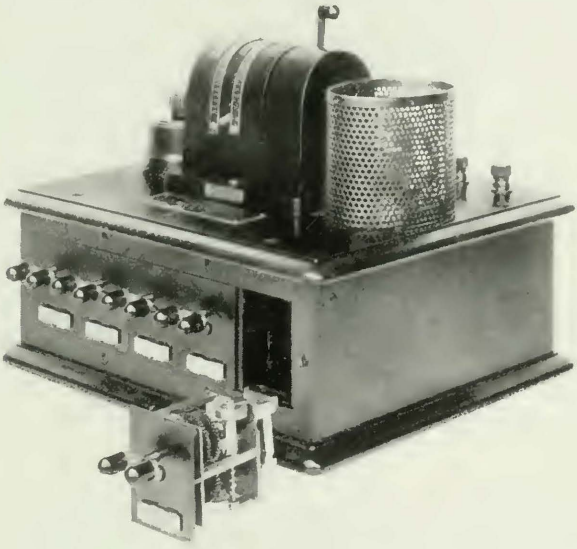


Fig.1—Transmitter attachment

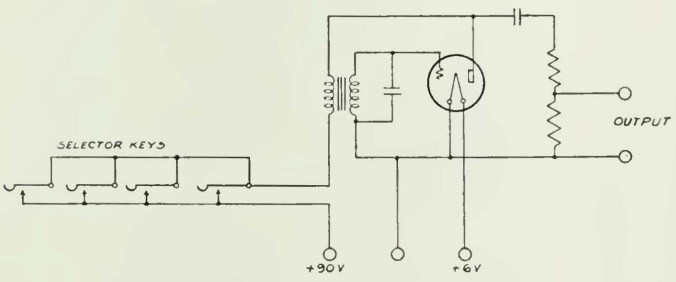


Fig. 2—Schematic of transmitter attachment

keys, which are connected in parallel, are in series with the plate winding of the oscillator coil. Operating any one of the calling keys opens and closes its contacts in a regular sequence determined by the code for which the key is set. Five group keys and a master key are

provided. Each of the group keys is set to a certain code call which may be changed by resetting the key by means of a screw-driver. One of these group keys may be used as a call for the entire system so that by the operation of this key, every receiver is called simultaneously. The other keys may be used for the four main group calls. The master key is similar in appearance to a miniature cash register and by setting its levers to the proper code combination, any desired station may be called individually and this key may also be used for the sub-group combinations.

The output terminals of the oscillator of the transmitter attachment may be connected directly to the speech input equipment of a standard radio telephone transmitter in place of the microphone. The speech input amplifier is adjusted so that when the signaling attachment is used the maximum possible modulation is obtained. The sensitivity of the radio receiver is adjusted for reliable operation of the signaling system, which is sharply tuned to the frequency of the transmitter attachment. This frequency is 3,000 cycles and is so high that the volume of music or speech will be amply sufficient for easy reception but yet insufficient to operate the signaling system relays as only a relatively small proportion of the energy in normal speech or music occurs at frequencies in the vicinity of 3,000 cycles. Even should the relays be operated occasionally by excessive volume of speech or music the receiver signal lamp will not light unless the proper code call is sent.

Receiving Apparatus

Photographs of the receiving equipment are shown in Figs. 3, 4 and 5. At some of the outlying stations of the New York Police Department 110-volt DC power supply is available while at others 110 volt 60 cycle AC is provided. The radio receivers are made in two different types, one type for each kind of power supply. A schematic circuit of the DC type of receiver is shown in Fig. 6 and that of the AC type in Fig. 7. These two types are similar in all respects except such modifications as the different sources of power supply necessitate.

In the direct current type of receiver all of the vacuum tube filaments are connected in series, current being taken directly from the line through a filter to eliminate line noises due to generator hum and other causes. In parallel with each filament is connected a small switchboard lamp with a red cap mounted on the panel of the receiver cabinet. The resistance of each lamp filament is sufficiently greater than that of the vacuum tube filament so that the lamp will light

only when the vacuum tube filament burns out or the tube is removed from its socket. In order to indicate when the power is turned on the receiver there is another lamp (green) mounted on the panel and connected across the 110 volt direct current line on the receiver side of the main switch so that when the switch is closed the lamp will light.

In the alternating current type of receiver the filaments of the



Fig. 3—Exterior view of receiving set. (The external appearance of the A.C. and D.C. models is the same)

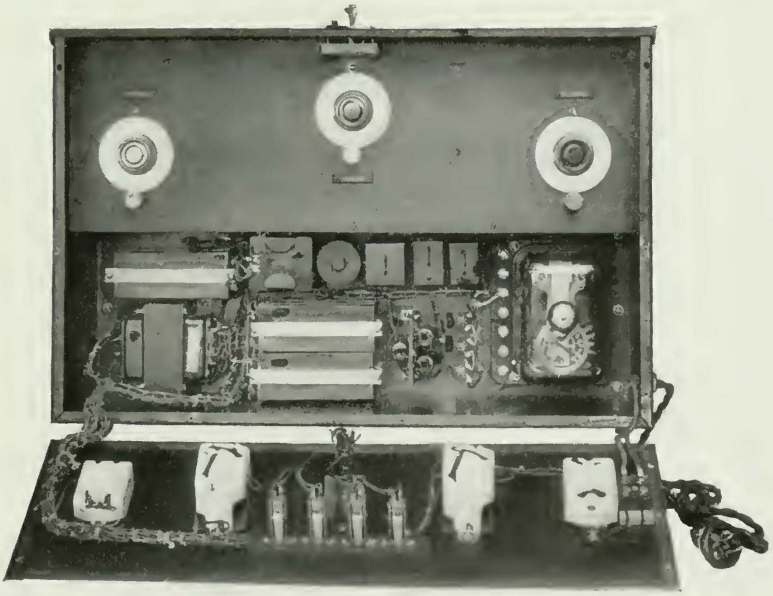


Fig. 4—View of A.C. model receiver showing tuning controls, and selector, and rectifying apparatus

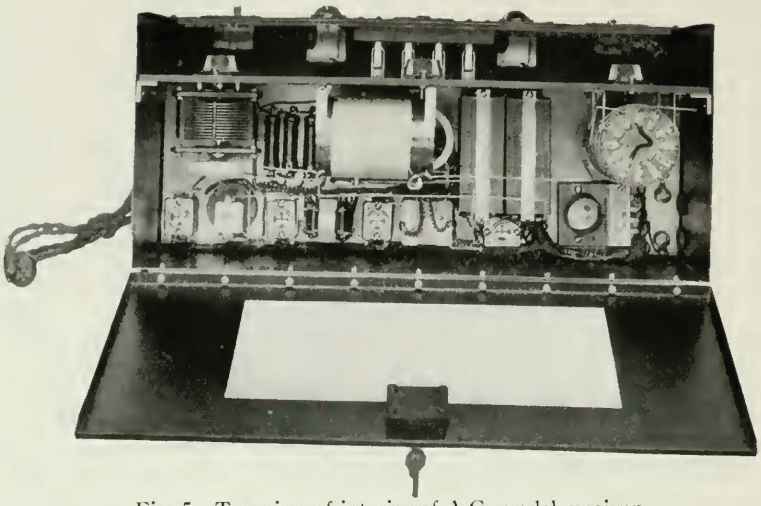


Fig. 5—Top view of interior of A.C. model receiver

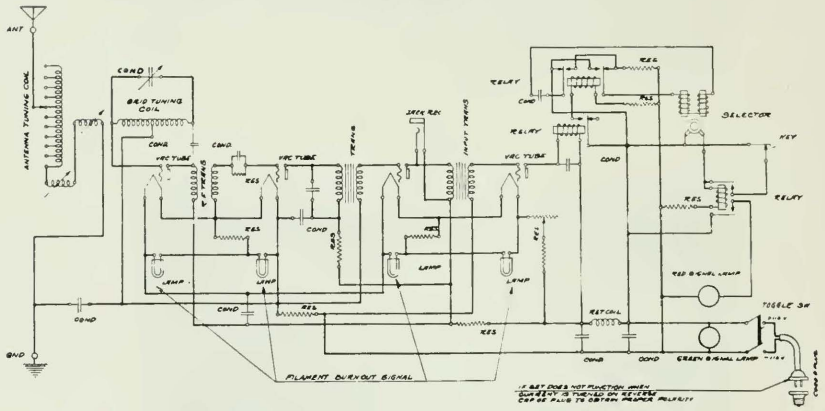


Fig. 6—Schematic of D.C. receiver

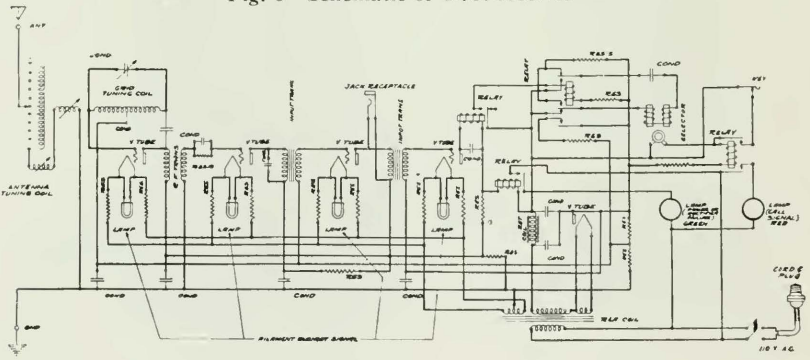


Fig. 7—Schematic of A.C. receiver

vacuum tubes are connected in parallel and are operated on alternating current. In order to take care of the filament warning lamps a resistance is provided in series with each tube filament so that when this filament burns out the voltage will rise sufficiently to light the corresponding lamp. To supply the necessary direct current for the plate potential of the vacuum tubes and for the operation of the relays, this receiver is provided with a power rectifier tube. The rectifier tube circuit is connected to the high voltage secondary winding of a power transformer and in connection with a filter system supplies to the receiver 110 volts direct current. As an indication of a burnt out rectifier tube filament there is provided a relay connected in series with the rectified current and a green signal lamp connected across the line in series with the contacts of this relay. The lamp will light when the power is turned on only if the rectifier tube is functioning properly. Thus it serves two useful purposes.

Both the DC and AC types of radio receivers are designed for operation from an open antenna, a double tuned, inductively coupled circuit being employed. The antenna circuit is tuned by means of taps on the loading coil and a small series inductance whose coupling to the loading coil is variable. An adjustable coupling coil is also included in the antenna circuit and serves as a sensitivity control. The secondary circuit to which this coil is coupled is tuned by means of a variable air condenser. All of these controls are on a panel inside the receiver and are inaccessible when the receiver cabinet is locked, thus insuring their remaining in adjustment.

Four "peanut" vacuum tubes are employed in the radio receiver. The first tube serves as a radio frequency amplifier. By means of a fixed radio frequency transformer sharply tuned to the frequency of the transmitting station, this tube is coupled to the second tube, which acts as a grid leak detector. This arrangement provides additional selectivity and more amplification than if a broadly tuned transformer were used, and is permissible since, in any given system, it is anticipated that the transmitter frequency will remain constant. To adapt the receiver to operate at any other transmitting frequency the transformers may be readily replaced by others of the proper characteristics. The third tube serves as an audio-frequency amplifier, being coupled to the detector tube by means of an audio-frequency transformer having a frequency characteristic suitable for the transmission of speech or music. The fourth tube serves as a rectifier and is coupled to the amplifier by means of a transformer, sharply tuned to the signaling frequency of 3,000 cycles. This frequency, as noted above, is sufficiently above the preponderant frequencies

of normal speech or music to make the accidental operation of the signaling system a very remote possibility.

The normal plate current in the rectifier tube is a small fraction of a milliampere. Upon a signal being received, this current increases sufficiently to operate a relay connected in the plate circuit. The operation of this relay closes a circuit through the winding of another relay, which is the reversing relay for the operation of the selector controlling the red signal lamp which indicates a call.

The selector is the heart of the signaling system. This selector is shown in Fig. 8. It consists of a mechanism unit mounted upon a magnet unit, the whole being enclosed in a glass case for protection.

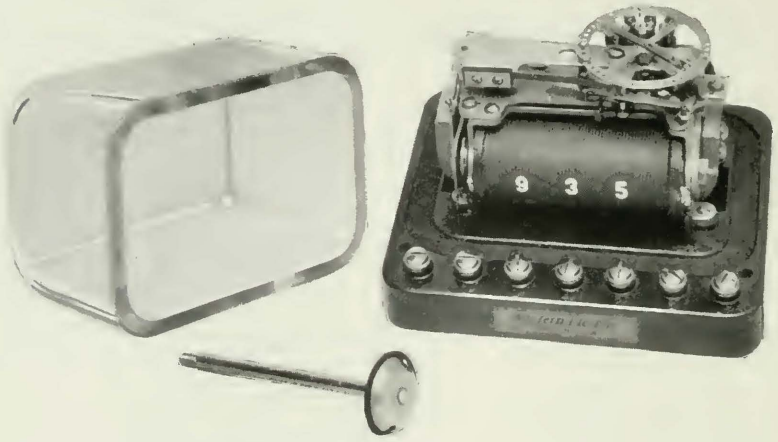


Fig. 8—Selector

The magnet windings are connected to a source of direct current of suitable voltage in series with a large fixed condenser and the contacts of the reversing relay mentioned above. When this relay is in either end position the circuit is completed through the condenser and the windings of the selector. As the power source is direct current it serves only to charge the condenser to its own potential. When the reversing relay is alternately operated and released, however, the repeated charging of this condenser in opposite directions sends pulses of current through the windings of the selector. This gives a rocking motion to the armature of the selector which motion is transmitted through a ratchet to the selector code wheel, advancing it against the tension of a spring one step for each movement of the armature. If the code wheel is kept from flying back during the pauses, two in number, which occur between groups of impulses, as is the case when the signal

corresponds to the code setting of a particular selector, it will be advanced step by step until it reaches the point at which it makes signaling contact. For any other signal, however, it will be released at the pause between some two groups of pulses and will then immediately fly back to its initial position. To hold the code wheel during signal pauses a series of pins is arranged to engage a spring arm on the selector frame, their position thus determining the code of the particular selector. As the master key can be so operated as to send signals without any pause a general call can be made through all selectors simultaneously whatever their code setting. When the code wheel has rotated a distance corresponding to twenty-seven impulses, a spring contact mounted upon it makes contact with the first of a series of four stationary contacts A, B, C, and D. Two more impulses make contact with the second of the series, two more with the third, and two more, or a total of 33, with the fourth. Only one of these contacts is connected in any individual selector and four large groups are thus provided. All the transmitting keys are so arranged that about one second after the completion of their calling signal, they send a signal which restores all selectors under their control to normal.

By omitting one or both of the pauses between the three groups of impulses, it is apparent that each selector will respond to four and only four systems of pulses for each contact. On contact A, for example, the selector will close the signaling circuit if its individual call is sent, if the first pause only or the second pause only is omitted, or if both pauses are omitted and 27 consecutive impulses are sent. All of the selectors using any one of the four contacts may thus be grouped in several different ways. The total possible number of individual stations in each of the four large groups is somewhat over 200, or over 800 in the entire system, the exact number depending upon the grouping system which is employed. Each large group may be subdivided into a number of small groups having from 15 to 20 stations in each group, of which each station may belong to two groups if desired. In any case the number of consecutive impulses without pauses corresponding to the contact used will call all the stations in that large group, the sub-groups being formed by omitting one of the pauses. The system is thus capable of a high degree of flexibility.

The operation of the selector closes a circuit through the winding of a relay. This relay is of the slow operating type, this being necessary in order to avoid signals due to momentary contacts which are made by the selector with certain code combinations. The relay is

provided with a holding contact so that after the selector has been restored to normal by the releasing impulse, it will remain operated until the person at the receiving station presses the key which opens the circuit of the holding contact. In the operated position, the relay completes the circuit of the red signal lamp on the panel of the receiver, thus indicating to the operator that he has been called. An audible signal may also be connected in parallel with it, using an additional relay if necessary to handle the heavy current required by a large gong.

In the exterior views of the receivers at the extreme left of the receiver is shown a jack into which the head telephones may be plugged. Plugging in these telephones does not interfere in any way with the operation of the signal system. Just to the right of the telephone jack is shown the red signal lamp for indicating when the receiver has been called. In the middle of the panel is a key which is used by the operator of the receiver to extinguish the red signal lamp when he takes up his head telephones. To the right of the key is the green signal lamp indicating when the power is on the receiver, and to the extreme right is the switch for turning the power on and off.

Wave Propagation in Overhead Wires with Ground Return

By JOHN R. CARSON

I

THE problem of wave propagation along a transmission system composed of an overhead wire parallel to the (plane) surface of the earth, in spite of its great technical importance, does not appear to have been satisfactorily solved.[1] While a complete solution of the actual problem is impossible, on account of the inequalities in the earth's surface and its lack of conductive homogeneity, the solution of the problem, where the actual earth is replaced by a plane homogeneous semi-infinite solid, is of considerable theoretical and practical interest. The solution of this problem is given in the present paper, together with formulas for calculating inductive disturbances in neighboring transmission systems.

The axis of the wire is taken parallel to the z-axis at height h above the xz-plane and passes through the y-axis at point O' as shown in Fig. 1 herewith. The "image" of the wire is designated by O''.

For $y>0$ (in the dielectric) the medium is supposed to have zero conductivity, while for $y<0$ (in the ground) the conductivity of the medium is designated by λ. The xz-plane represents the surface of separation between dielectric and ground.

We consider a wave propagated along the z-axis and the current, charge and field are supposed to contain the common factor $\exp(-\Gamma z+i\omega t)$, which, however, will be omitted for convenience in the formulas. The propagation constant, Γ, is to be determined. It is assumed, *ab initio*, as a very small quantity in c.g.s. units.[2]

In the ground ($y\leqq 0$) the axial electric force is formulated as the

[1] See Rudenberg, Zt. f. Angewandt, Math. u. Mechanik, Band 5, 1925. In that paper the current density in the ground is assumed to be distributed with radial symmetry. The resulting formulas are not in agreement with those of the present paper. Since this paper was set up in type I have learned that formulas equivalent to equations (26), (28), (31) for the *mutual impedance* of two parellel wires were obtained by my colleague, Dr. G. A. Campbell, in 1917. It is to be hoped that his solution will be published shortly.

[2] The simplifying assumptions introduced in this analysis are essentially the same as those employed and discussed in "Wave Propagation Over Parallel Wires: The Proximity Effect," *Phil. Mag.*, Vol. xli, April, 1921.

general solution, symmetrical with respect to x, of the wave equation; thus

$$E_z = -\int_0^\infty F(\mu) \cos x\mu \; e^{y\sqrt{\mu^2+i\alpha}} d\mu, \qquad y \leqq 0 \tag{1}$$

where

$\alpha = 4\pi\lambda\omega$,

λ = conductivity of ground in elm. c.g.s. units,

$i = \sqrt{-1}$,

$\omega/2\pi$ = frequency in cycles per second.

(In the following analysis and formulas, elm. c.g.s. units are employed throughout).

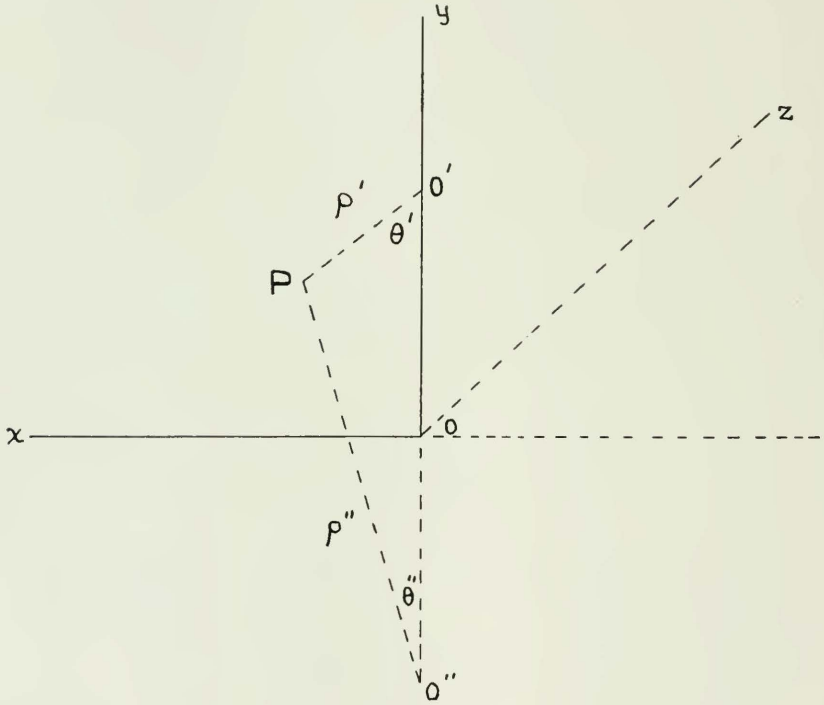


Fig. 1

Assuming that in the ground E_x and E_y are negligible compared with E_z, we have from the formula, curl $E = -\frac{\partial}{\partial t} H$,

$$i\omega H_x = -\frac{\partial}{\partial y} E_z$$

$$i\omega H_y = \frac{\partial}{\partial x} E_z.$$

Whence, in the ground

$$H_x = \frac{1}{i\omega}\int_0^\infty \sqrt{\mu^2+i\alpha}\,.F(\mu).\cos x\mu.e^{y\sqrt{\mu^2+i\alpha}}d\mu, \tag{2}$$

$$H_y = \frac{1}{i\omega}\int_0^\infty \mu.F(\mu).\sin x\mu.e^{y\sqrt{\mu^2+i\alpha}}d\mu, \tag{3}$$

it being understood that $y \leqq 0$. The function $F(\mu)$ in the preceding formulas is to be determined by the boundary conditions.

In the dielectric, H_x and H_y may be regarded as composed of two components; thus

$$H_x = H_x^0 + H_x',$$
$$H_y = H_y^0 + H_y',$$

where H_x^0, H_y^0 designate the field due to the current I in the wire, and H_x', H_y' the field of the ground current.

Neglecting axial displacement currents in the dielectric, and assuming that the wire is of sufficiently small radius so that the distribution of current over its cross section is symmetrical, we have

$$H_x^0 = \frac{\cos\Theta'}{\rho'}.\,2I, \tag{4}$$

$$H_y^0 = \frac{\sin\Theta'}{\rho'}.\,2I, \tag{5}$$

where $$\rho' = \sqrt{x^2+(y-h)^2},$$

$$\cos\Theta' = \frac{h-y}{\rho'}, \tag{6}$$

$$\sin\Theta' = x/\rho'.$$

The secondary magnetic field H_x', H_y' is taken as

$$H_x' = \int_0^\infty \phi(\mu)\cos x\mu.e^{-y\mu}d\mu, \tag{7}$$

$$H_y' = -\int_0^\infty \phi(\mu)\sin x\mu.e^{-y\mu}d\mu. \tag{8}$$

At the surface of separation $y=0$, H_x^0, H_y^0 are expressible as the Fourier integrals

$$H_x^0 = 2I\int_0^\infty \cos x\mu.e^{-h\mu}d\mu, \tag{9}$$

$$H_y^0 = 2I\int_0^\infty \sin x\mu.e^{-h\mu}d\mu. \tag{10}$$

Also at the surface of separation of the two media ($y=0$), H_x and H_y must be continuous. Equating the values of H_x and H_y at $y=0$, as given by (2), (3) and by (7), (8) and (9), (10), we have

$$\frac{1}{i\omega}\sqrt{\mu^2+i\alpha}.F(\mu)=2I.e^{-h\mu}+\phi(\mu),$$

$$\frac{1}{i\omega}\mu.F(\mu)=2I.e^{-h\mu}-\phi(\mu),$$

whence

$$F(\mu)=\frac{i\omega e^{-h\mu}}{\sqrt{\mu^2+i\alpha}+\mu}4I, \tag{11}$$

$$\phi(\mu)=\frac{(\sqrt{\mu^2+i\alpha}-\mu)}{\sqrt{\mu^2+i\alpha}+\mu}e^{-h\mu}.2I, \tag{12}$$

which determines the functions $F(\mu)$ and $\phi(\mu)$.

Inserting the value of $F(\mu)$, as given by (11) in (1), the axial electric force E_z in the ground ($y\leqq 0$) and therefore the distribution of current density in the ground is expressed as a Fourier integral in terms of the frequency $\omega/2\pi$, the current I in the wire, the height h of the wire above ground, and the conductivity λ of the ground. Similarly the insertion of $\phi(\mu)$, as given by (12) in formulas (7) and (8) gives the magnetic field H_x, H_y in the dielectric. Thus

$$E_z(x,y)=E_z=-i4\omega I\int_0^\infty \frac{e^{-\mu h}}{\sqrt{\mu^2+i\alpha}+\mu}e^{y\sqrt{\mu^2+i\alpha}}\cos x\mu.d\mu, \qquad y\leqq 0. \tag{13}$$

This can be further simplified if we write

$$x'=x\sqrt{\alpha}$$

$$y'=y\sqrt{\alpha}$$

$$h'=h\sqrt{\alpha},$$

whence

$$E_z=-4\omega I\int_0^\infty (\sqrt{\mu^2+i}-\mu)e^{-h'\mu}.e^{y'\sqrt{\mu^2+i}}\cos x'\mu\, d\mu, \qquad y\leqq 0. \tag{14}$$

The axial electric force in the dielectric is now to be formulated. This is always derivable from a vector and a scalar potential; thus

$$E_z=-i\omega A_z-\frac{\partial}{\partial z}V, \tag{15}$$

where A_z is the vector potential of the axial currents, and V the scalar potential. Consequently,

$$E_z(x,y)-E_z(x,0)=-i\omega(A_z(x,y)-A_z(x,0))-\frac{\partial}{\partial z}(V(x,y)-V_0). \tag{16}$$

Here $E_z(x,0)$ is the axial electric intensity at the surface of the ground plane ($y=0$), and

$$A_z(x,y)-A_z(x,0)=\int_0^y H_x(x,y)dy. \tag{17}$$

$V(x,y)-V_0$ is the difference in the scalar potential between the point x,y and the ground, which is due to the charges on the wire and on the surface of the ground. For convenience, it will be designated by V.

By means of (16) and the preceding formulas we get [3]

$$\begin{aligned} E_z=&-4\omega I\int_0^\infty(\sqrt{\mu^2+i}-\mu)e^{-(h'+y')\mu}\cos x'\mu\, d\mu \\ &-i\,2\omega I\log(\rho''/\rho')-\frac{\partial}{\partial z}V, \qquad y\geqq 0 \end{aligned} \tag{18}$$

where

$$\begin{aligned} \rho' &= \sqrt{(h-y)^2+x^2} \\ &= \text{distance of point } x,y \text{ from wire,} \\ \rho'' &= \sqrt{(h+y)^2+x^2} \\ &= \text{distance of point } x,y \text{ from image of wire.} \end{aligned}$$

The first two terms on the right hand side of (18) represent the electric force due to the varying magnetic field; the term $-\frac{\partial}{\partial z}V$ represents the axial electric intensity due to the charges on the surface of the wire and the ground. If Q be the charge per unit length, V is calculable by usual electrostatic methods on the assumption that the surface of the wire and the surface of the ground are equipotential surfaces, and their difference of potential is Q/C where C is the electrostatic capacity between wire and ground.[4]

II

By aid of the preceding analysis and formulas, we are now in a position to derive the propagation constant, Γ, and characteristic impedance, K, which characterize wave propagation along the system. Let z denote the "internal" or "intrinsic" impedance of the wire per

[3] As a check on this formula note that together with (14) it satisfies the condition of continuity of E_z at $y=0$.

[4] See "Wave Propagation Over Parallel Wires: The Proximity Effect," *Phil. Mag.*, Vol. xli, Apr., 1921.

unit length. (With small error this may usually be taken as the resistance per unit length of the wire.) The axial electric intensity at the surface of the wire is then zI. Equating this to the axial electric intensity at the surface of the wire as given by (18) and replacing $\partial/\partial z$ by $-\Gamma$, we have

$$\begin{aligned} zI = &-4\omega I \int_0^\infty (\sqrt{\mu^2+i}-\mu)e^{-2h'\mu}d\mu \\ &-i\,2\omega I.\log(\rho''/a)+\Gamma V. \end{aligned} \tag{19}$$

Writing $V=Q/C$ and

$$i\omega Q=\Gamma I-GV=\Gamma I-\frac{G}{C}Q,$$

where G is the leakage conductance to ground per unit length, we have, solving for Γ,

$$\Gamma^2=(G+i\omega C)[z+i2\omega.\log(\rho''/a)]+4\omega\int_0^\infty(\sqrt{\mu^2+i}-\mu)e^{-2h'\mu}d\mu). \tag{20}$$

Writing this in the usual form

$$\Gamma^2=(R+iX)(G+i\omega C), \tag{21}$$

the characteristic impedance is given by

$$K^2=\frac{R+iX}{G+i\omega C} \tag{22}$$

and the *series impedance per unit length of the circuit* is

$$R+iX=Z=z+i2\omega.\log(\rho''/a)+4\omega\int_0^\infty(\sqrt{\mu^2+i}-\mu)e^{-2h'\mu}d\mu. \tag{23}$$

It will be observed that the first two terms on the right hand side of (23) represent the series impedance of the circuit *if the ground is a perfect conductor*; the infinite integral formulates the effect of the finite conductivity of the ground.

The *mutual impedance*[5] Z_{12} between two parallel ground return circuits with wires at heights h_1 and h_2 above ground and a separation x between their vertical planes is given by

$$Z_{12}=i2\omega.\log(\rho''/\rho')+4\omega\int_0^\infty(\sqrt{\mu^2+i}-\mu)e^{-(h_1'+h_2')\mu}\cos x'\mu\,d\mu, \tag{24}$$

[5] It will be noted that the mutual impedance is equal to the axial electric intensity at the axis of the second wire due to the varying magnetic field of unit current in the first wire and its accompanying distribution of ground current.

where

$$\rho'' = \sqrt{(h_1+h_2)^2+x^2}$$

$$\rho' = \sqrt{(h_1-h_2)^2+x^2}$$

$$h_1' = h_1\sqrt{\alpha}$$

$$h_2' = h_2\sqrt{\alpha}$$

$$x' = x\sqrt{\alpha}.$$

From the preceding the series *self impedance* of the ground return circuit may be conveniently written as

$$Z = Z^0 + Z' \tag{25}$$

and the *mutual impedance* as

$$Z_{12} = Z_{12}^0 + Z_{12}' \tag{26}$$

where Z^0, Z_{12}^0 are the self and mutual impedances respectively, *on the assumption of a perfectly conducting ground*, and

$$Z' = 4\omega \int_0^\infty (\sqrt{\mu^2+i}-\mu)e^{-2h'\mu}d\mu, \tag{27}$$

$$Z_{12}' = 4\omega \int_0^\infty (\sqrt{\mu^2+i}-\mu)e^{-(h_1'+h_2')\mu} \cos x'\mu \, d\mu. \tag{28}$$

The calculation of the circuit constants and the electromagnetic field in the dielectric depends, therefore, on the evaluation of an infinite integral of the form

$$J(p,q) = J = \int_0^\infty (\sqrt{\mu^2+i}-\mu)e^{-p\mu}. \cos q\mu \, d\mu. \tag{29}$$

In terms of this integral

$$Z' = 4\omega.J(2h'.0) \tag{30}$$

$$Z_{12}' = 4\omega.J(h_1'+h_2', x'). \tag{31}$$

To the solution of the infinite integral $J(p,q)$ we now proceed.

III

The solution of equation (29), that is, the evaluation of $J(p,q)$ can be made to depend on the solution of the infinite integral

$$\int_0^\infty \sqrt{\mu^2+\alpha^2}.e^{-\beta\mu}d\mu$$

which has been worked out and communicated to me by R. M. Foster. It is

$$\frac{\alpha}{\beta}\left\{ K_1(\alpha\beta)+G(\alpha\beta) \right\}$$

where $K_1(x)$ is the Bessel function of the second kind and first order as defined by Jahnke und Emde, Funktionentafeln, pg. 93, and $G(x)$ is the absolutely convergent series

$$G(x)=\frac{x^2}{3}-\frac{x^4}{3^2.5}+\frac{x^6}{3^2.5^2.7}-\ \ldots\ .$$

On the basis of this solution, it is a straightforward though intricate and tedious process to derive the following solution for $J(p,q)$ of equation (29).

Writing $r=\sqrt{p^2+q^2}$ and $\Theta=\tan^{-1}(q/p)$, it is $J=P+iQ$

in which

$$\begin{aligned} P=&\frac{\pi}{8}(1-s_4)+\frac{1}{2}\left(\log\frac{2}{\gamma r}\right)s_2+\frac{1}{2}\Theta.s_2' \\ &-\frac{1}{\sqrt{2}}\sigma_1+\frac{1}{2}\sigma_2+\frac{1}{\sqrt{2}}\sigma_3, \end{aligned} \tag{32}$$

$$\begin{aligned} Q=&\frac{1}{4}+\frac{1}{2}\left(\log\frac{2}{\gamma r}\right)(1-s_4)-\frac{1}{2}\Theta.s_4' \\ &+\frac{1}{\sqrt{2}}\sigma_1-\frac{\pi}{8}s_2+\frac{1}{\sqrt{2}}\sigma_3-\frac{1}{2}\sigma_4. \end{aligned} \tag{33}$$

In these equations $\log\gamma$ is Euler's constant:

$\gamma=1.7811$, $\log\frac{2}{\gamma}=0.11593$, $\log\gamma=0.57722$ and σ_1, σ_2, σ_3, σ_4, s_2, s_2', s_4, s_4', are infinite series defined as follows:

$$s_2=\frac{1}{1!2!}\left(\frac{r}{2}\right)^2\cos 2\Theta-\frac{1}{3!4!}\left(\frac{r}{2}\right)^6\cos 6\Theta+\ \ldots\ ,$$

$$s_2'=\frac{1}{1!2!}\left(\frac{r}{2}\right)^2\sin 2\Theta-\frac{1}{3!4!}\left(\frac{r}{2}\right)^6\sin 6\Theta+\ \ldots\ ,$$

$$s_4=\frac{1}{2!3!}\left(\frac{r}{2}\right)^4\cos 4\Theta-\frac{1}{4!5!}\left(\frac{r}{2}\right)^8\cos 8\Theta+\ \ldots$$

$$s_4' = \frac{1}{2!3!}\left(\frac{r}{2}\right)^4 \sin 4\theta - \frac{1}{4!5!}\left(\frac{r}{2}\right)^8 \sin 8\theta + \ldots,$$

$$\sigma_1 = \frac{r\cos\theta}{3} - \frac{r^5\cos 5\theta}{3^2.5^2.7} + \frac{r^9\cos 9\theta}{3^2.5^2.7^2.9^2.11} - \ldots,$$

$$\sigma_3 = \frac{r^3\cos 3\theta}{3^2.5} - \frac{r^7\cos 7\theta}{3^2.5^2.7^2.9} + \frac{r^{11}\cos 11\theta}{3^2.5^2.7^2.9^2.11^2.13} - \ldots,$$

$$\sigma_2 = \left(1+\frac{1}{2}-\frac{1}{4}\right)\frac{1}{1!2!}\left(\frac{r}{2}\right)^2 \cos 2\theta$$

$$-\left(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}-\frac{1}{8}\right)\frac{1}{3!4!}\left(\frac{r}{2}\right)^6 \cos 6\theta + \ldots,$$

$$= \frac{5}{4}s_2 \text{ approximately,}$$

$$\sigma_4 = \left(1+\frac{1}{2}+\frac{1}{3}-\frac{1}{6}\right)\frac{1}{2!3!}\left(\frac{r}{2}\right)^4 \cos 4\theta$$

$$-\left(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}-\frac{1}{10}\right)\frac{1}{4!5!}\left(\frac{r}{2}\right)^8 \cos 8\theta + \ldots.$$

$$= \frac{5}{3}s_4 \text{ approximately.}$$

It is to be regretted that the foregoing formulas appear so complicated. The series, however, are very rapidly convergent and for $r \leqq 2$ only the two leading terms of each series need be retained. For $r \leqq 1$, only the leading terms are of importance.

For the important range $r \leqq 1/4$,

$$P = \frac{\pi}{8} - \frac{1}{3\sqrt{2}} r\cos\theta + \frac{r^2}{16}\cos 2\theta\left(0.6728 + \log\frac{2}{r}\right) + \frac{r^2}{16}\theta\sin 2\theta, \quad (34)$$

$$Q = -0.0386 + \frac{1}{2}\log\left(\frac{2}{r}\right) + \frac{1}{3\sqrt{2}} r\cos\theta. \quad (35)$$

For $r > 5$ the following asymptotic expansions, derivable from (29) by repeated partial integrations, are to be employed.

$$P = \frac{1}{\sqrt{2}}\frac{\cos\theta}{r} - \frac{\cos 2\theta}{r^2} + \frac{1}{\sqrt{2}}\frac{\cos 3\theta}{r^3} + \frac{3}{\sqrt{2}}\frac{\cos 5\theta}{r^5} \ldots, \quad (36)$$

$$Q = \frac{1}{\sqrt{2}}\frac{\cos\theta}{r} - \frac{1}{\sqrt{2}}\frac{\cos 3\theta}{r^3} + \frac{3}{\sqrt{2}}\frac{\cos 5\theta}{r^5} - \ldots, \quad (37)$$

$r > 5.$

For large values of $r(r>10)$, these reduce to

$$J=\frac{1+i}{\sqrt{2}}\frac{\cos\Theta}{r}-\frac{\cos 2\Theta}{r^2},\ r>10. \tag{38}$$

In view of the somewhat complicated character of the function in the range $1/4\leqq r\leqq 5$ the curves shown below have been computed. These show $J=P+iQ$ as a function of r for $\Theta=0, \frac{\pi}{8}, \frac{\pi}{4}, \frac{3\pi}{8}, \frac{\pi}{2}$. By interpolation it is possible to estimate with fair accuracy the value of the functions for intermediate values of Θ.

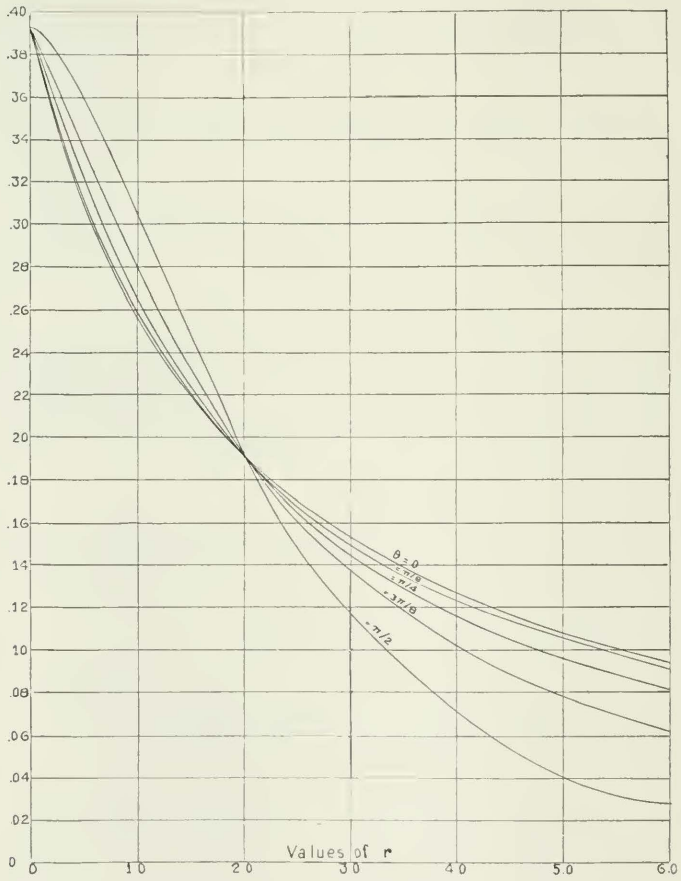


Fig. 2 P = real part of J

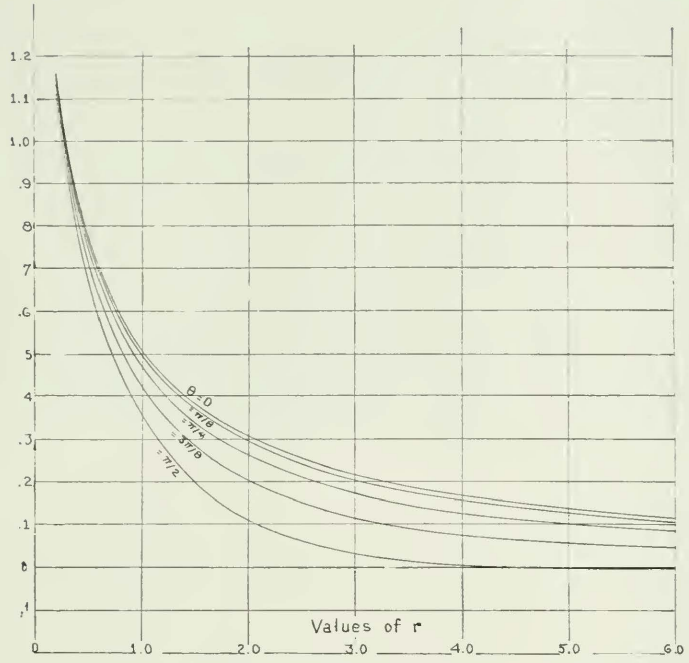


Fig. 3 Q = imaginary part of J

IV

The theory and formulas of the preceding sections will now be reviewed and summarized as regards their principal applications to technical transmission problems where the ground forms, in whole or part, the "return" part of the circuit. In such problems we are interested in the electric intensity in the dielectric and in the ground, and in particular in the self impedance and mutual impedances of ground return circuits. The calculation of these quantities is provided for by the general analysis and the solution of the infinite integral J. Reference should be made to Fig. 1 shown in section I for the geometry of the system and coordinate system employed.

1. *The Axial Electric Intensity E_z in the Dielectric.* (See equations (15) and (18)).

$$E_z = -\frac{\partial}{\partial z} V - (i2\omega \log (\rho''/\rho') + 4\omega J) I$$

where

$$\rho' = \sqrt{(h-y)^2+x^2}$$

$$= \text{distance of point in dielectric } \textit{from axis} \text{ of wire.}$$

$$\rho'' = \sqrt{(h+y)^2+x^2}$$

$$= \text{distance of point in dielectric } \textit{from image} \text{ of wire.}$$

$$r = \rho''\sqrt{\alpha}$$

$$\Theta = \sin^{-1}(x/\rho'')$$

$$\alpha = 4\pi\lambda\omega.$$

These values of r and Θ are, of course, to be employed in calculating $J = P + iQ$ from the formulas and curves of the preceding section. As a special case *the electric intensity at the surface of the earth is*

$$E_z = -4\omega J I$$

$$\rho'' = \sqrt{h^2+x^2}$$

$$r = \rho''\sqrt{\alpha}$$

$$\Theta = \sin^{-1}(x/\rho'')$$

2. *Self Impedance of Ground Return Circuit.* (See equations (25), (27), (30)).

$$Z = Z^0 + 4\omega J$$

Z^0 = self impedance with perfectly conducting ground.

$$r = 2h\sqrt{\alpha}$$

$$\Theta = 0.$$

3. *Mutual Impedance of Ground Return Circuits.* (See equations (26), (28), (31)).

$$Z_{12} = Z_{12}^0 + 4\omega J$$

Z_{12}^0 = mutual impedance with perfectly conducting ground.

$$r = \sqrt{\alpha}\sqrt{(h_1+h_2)^2+x^2} = \rho''\sqrt{\alpha}$$

$$\Theta = \sin^{-1}(x/\rho'').$$

The axial electric intensity E_z in the ground ($y<0$) is given by equation (1), and the subsequent analysis, whence

$$E_z = -4\omega I \int_0^\infty (\sqrt{\mu^2+i}-\mu).\cos x'\mu.e^{-(h'\mu+g'\sqrt{\mu^2+i})}\,d\mu$$

where, as before

$$x' = x\sqrt{\alpha}$$

$$h' = h\sqrt{\alpha}$$

and

$g' = \sqrt{\alpha}$ times the depth below the surface of the ground.

$$= g\sqrt{\alpha}.$$

The integral can undoubtedly be evaluated in somewhat the same way as (29) and can in any case be numerically computed without much difficulty. Owing, however, to the secondary technical interest in the electric intensity below the surface of the earth, the detailed solution has not been undertaken, nor has the magnetic field been worked out.

V

The practical utility of the preceding theory and formulas will now be illustrated by a brief sketch of their application to two important transmission problems.

The Wave Antenna

When a transmission line with "ground return" is employed as a radio receiving antenna it is called a wave antenna. The theory and design of such an antenna requires a knowledge of the transmission characteristics of the ground return circuit, which are calculable, as shown above, from the geometry and constants of the overhead wire, together with $Z' = 4\omega J$, which may be termed the "ground return" impedance.

We assume that the wire is approximately 30 ft. above the ground ($h = 10^3$) and that the frequency is 5.10^4 c.p.s. corresponding to the frequency employed in Trans-Atlantic radio communication. The ground conductivity λ is exceedingly variable, depending on the locality and weather conditions. Calculations of Z' will therefore be made for two extreme cases, $\lambda = 10^{-12}$ and $\lambda = 10^{-14}$ which should cover the range of variation encountered in practice.

For $\lambda = 10^{-12}$,

$$\sqrt{\alpha} = \sqrt{4\pi\lambda\omega} = 2.10^{-3}$$

and for $\lambda = 10^{-14}$,

$$\alpha = 2.10^{-4}.$$

Correspondingly, $r=2h\sqrt{\alpha}$ has the values 4.0 and 0.4, respectively. Reference to the preceding formulas and curves for J, for $r=4.0$ and $r=0.4$, give

$$J=0.126+i\ 0.168, \qquad \lambda=10^{-12}$$

$$J=0.323+i\ 0.871, \qquad \lambda=10^{-14}$$

whence the corresponding values of Z' are

$$Z'=4\omega.\ (0.126+i\ 0.168),$$

$$Z'=4\omega.\ (0.323+i\ 0.871).$$

These are the "ground return" impedances per unit length in elm. c.g.s. units; to convert to *ohms per mile* they are to be multiplied by the factor 1.61×10^{-4}. Consequently setting $\omega=\pi.10^{5}$, we get

$$Z'=6.44\pi(1.3+i\ 1.7), \qquad \lambda=10^{-12}$$

$$Z'=6.44\pi\ (3.2+i\ 8.7), \qquad \lambda=10^{-14}.$$

Comparison of these formulas shows that an hundred-fold increase in the resistivity of the ground increases the resistance component of the ground return impedance by the factor 2.5 and increases its reactance only five-fold. That is to say, the ground return impedance is not sensitive to wide variations in the resistivity of the earth, a fortunate circumstance in view of its wide variability and our lack of precise information regarding it.

Induction from Electric Railway Systems

A particularly important application of the preceding analysis is to the problems connected with the disturbances induced in parallel communication lines by alternating current electric railways. Assuming the frequency as 25 c.p.s., we have corresponding to $\lambda=10^{-12}$ and $\lambda=10^{-14}$,

$$\sqrt{\alpha}=0.45\times10^{-4} \text{ and } 0.45\times10^{-5}.$$

Taking the height of the trolley wire as approximately 30 ft., $h=10^{3}$ and assuming the parallel telephone as the same height above ground and separated by approximately 120 ft., $x=4.10^{3}$, and

$$r=\sqrt{\alpha}\sqrt{(2h)^2-x^2}$$

$$=4.47\times10^{3}\sqrt{\alpha},$$

and corresponding to the values of α taken above

$$r = 0.2 \text{ and } 0.02 \text{ in round numbers,}$$

while

$$\Theta = \sin^{-1} \frac{4}{\sqrt{20}} = 63^\circ\ 30' \text{ approximately.}$$

For both cases, therefore, we can employ, in calculating $J = P + iQ$, the approximate formulas,

$$P = \frac{\pi}{8} - \frac{1}{3\sqrt{2}} r \cdot \cos \Theta + \frac{r^2}{16} \cos 2\Theta \left(.6728 + \log \frac{2}{r} \right) + \frac{r^2}{16} \Theta \sin 2\Theta,$$

$$Q = -0.0386 + \frac{1}{2} \log \left(\frac{2}{r} \right) + \frac{1}{3\sqrt{2}} r \cos \Theta.$$

For $\lambda = 10^{-12}$ and $r = 0.2$, this gives

$$J = 0.369 + i\ 1.135$$

and

$$Z'_{12} = 4\omega(0.369 + i\ 1.135).$$

The foregoing assumes that the only return conductor is the ground. If, however, an equal and opposite current flows in the rail we must subtract from the foregoing mutual impedance, the mutual impedance between rail and telephone line; that is, the mutual impedance Z'_2 between the telephone line and a conductor at the surface of the earth. For this case

$$\rho'' = \sqrt{h^2 + x^2} = 4.12 \times 10^3$$

$$\Theta = \sin^{-1} \frac{4}{\sqrt{17}} = 76^\circ$$

$$\cos \Theta = 0.242,\ r = 0.184 \text{ for } \lambda = 10^{-12}.$$

The corresponding value of J is

$$J = 0.378 + i\ 1.165$$

and the *resultant* mutual impedance between railway and parallel telephone line is,

$$Z'_{12} = 4\omega(0.369 - 0.378 + i\ (1.135 - 1.165))$$
$$= 4\omega(0.009 - i\ 0.030).$$

The very large reduction in mutual impedance, due to the current in the rail, is striking.

For the case of $\lambda = 10^{-14}$, the corresponding calculations give

$$Z'_{12} = 4\omega(0.391 + i\ 2.27)$$

with *no current in rail*, and

$$Z'_{12} = 4\omega(-0.001 - i\ 0.002)$$

with *equal and opposite current in rail*. It is evident from these figures that the reduction in mutual impedance, due to the current in the rail, is practically independent of the ground conductivity, at least at the separation specified.

Electrode Effects in the Measurement of Power Factor and Dielectric Constant of Sheet Insulating Materials

By E. T. HOCH

Synopsis: It is shown that, aside from the guard ring type of electrode which can only be used with certain special types of measuring circuit, the most accurate results can probably be obtained by the use of equal foil electrodes and making proper allowance for the edge effects. From the standpoint of convenience, mercury electrodes and foil electrodes of unequal size have certain advantages. It is believed that the results obtained with these two types are also sufficiently accurate for most purposes when the corresponding corrections are applied.

Introduction

WHEN it is desired to determine the dielectric constant and power factor of an insulating material, the first problem that presents itself is that of finding a suitable means of applying a potential to the material in question. In order to accomplish this, a sample of the material must, in general, be placed between two conductors or electrodes to which the desired potential is applied. The size, shape and manner of application of these electrodes affect directly the quantities to be measured from which the dielectric constant and power factor are derived. Therefore, unless proper allowance can be made for them, these features of the electrodes will affect the values obtained for the properties of the insulation.

It is the purpose of this paper to discuss only the part played by the electrodes in the measurement of power factor and dielectric constant and not to discuss complete methods for measuring these quantities. Experimental data will be presented to show the magnitude of the various effects discussed.

If any form of test is to be of general use for determining the properties of any material, there are certain fundamental requirements which must be fulfilled. First, the method should lead to exact reproducibility of results. That is, a test on a given sample of material should lead to the same result regardless of when, where or by whom the test is made. Second, the result obtained should be the correct result, that is, the absolute accuracy should be high. Third, the method should be as convenient as possible to use. If it is not, the method loses its practical value to a large extent since the tendency will be for most people to use a more convenient method even at the expense of accuracy and reproducibility.

Sources of Error

There are certain sources of error inherent in all electrodes which affect both the reproducibility and the accuracy of the results obtained. First, we have the question of contact between the electrode and the sample. If the electrode does not make perfect contact over its entire area with the sample, the result obtained is not a true value for the material of the sample, but is a resultant, depending upon the amount and nature of the material filling the gap. Air spaces between the electrode and the sample have a marked effect on the apparent dielectric constant. An air-gap. 001 in. thick in series with a sample having a dielectric constant of 5 will have the same effect on the capacitance as increasing the thickness of the sample by .005 in. If the actual thickness of the sample is .05 in. this results in an error of nearly 10% in the value of dielectric constant. The power factor will also be reduced and the loss factor or product of power factor and dielectric constant will be reduced by the factor $\left(\frac{50}{55}\right)^2$, or about 17%. Thus it is evident that the elimination of all gaps between the electrode and the sample is one of the first requirements both for reproducibility and accuracy.

A second effect inherent in all electrodes which is a source of error unless properly allowed for, is the so-called fringing of the electrostatic flux, that is, the lines of force tend to spread out and include an area of the sample greater than that of the electrode. This is

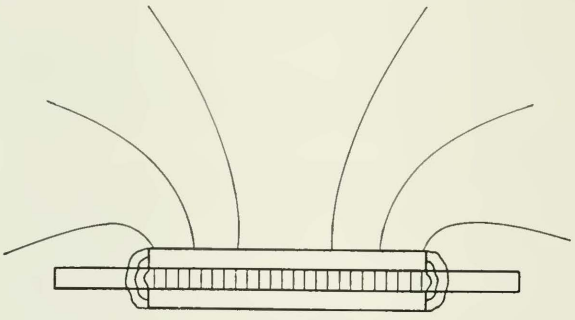


Fig. 1

illustrated in Fig. 1. So far as the flux which is confined entirely to the material of the sample is concerned, this produces an error in the dielectric constant, which involves a determination of the effective area of the sample, but not in the power factor which is independent of the area. However, there are some lines of force

which terminate on the vertical surfaces of the electrode and whose paths are partly through the sample and partly through air. These introduce a slight error into the power factor also. They also make the capacitance depend to a slight extent on the thickness of the electrodes. These edge effects vary with the thickness of the sample and also with the ratio of the perimeter to the area of the electrode and hence with its size and shape. They are also increased materially when one electrode is larger than the other.

The third inherent source of error is the capacitance from the ungrounded electrode [1] to earth due to lines of force which pass out in all directions other than through the sample. This also is illustrated in Fig. 1. This increases the measured capacitance by an amount depending somewhat on the nature and position of surrounding objects. If this capacitance is due to flux passing entirely through air it makes no difference in the dielectric loss,[2] but if the path of the flux includes other material such as the wood, brick, plaster, etc., in the walls and floor of the room as is often the case, it may add an appreciable loss.

We will now consider the above errors and also the question of convenience as applied to certain specific types of electrodes. The following types will be considered.

1. Plain metal electrodes.
2. Mercury electrodes.
   a—Confining ring of metal.
   b—Confining ring of insulating material.
3. Foil electrodes.
   a—Both same size as sample.
   b—Both same size, but smaller than the sample
   c—One materially larger than the other.
4. Conducting paint electrodes.
5. Fixed gap electrodes.

Plain Metal Electrodes

One of the simplest forms of electrode would be two similar blocks of metal between which the sample would be placed. If the surfaces of both the electrode and sample were true planes, this would be fairly satisfactory. However, as the surfaces of samples of insulating material usually available for test are not true planes, the air-gap

[1] Assuming one electrode to be grounded as is usually the case.

[2] The apparent power factor is reduced by the increase in capacitance but the loss factor is not affected since the dielectric constant is increased to the same extent that the power factor is reduced.

error is usually prohibitive and makes this type of electrode practically useless.

Sometimes electrodes of this type are amalgamated and flooded with mercury and then pressed onto the sample, the excess mercury being brushed away.

This is an improvement but still leaves considerable uncertainty as to the degree of contact obtained.

Mercury Electrodes

Primarily on account of the ease with which a liquid will conform to the contour of an irregular surface, mercury has frequently been used as an electrode material. The usual procedure is to float the sample in a tray of mercury which serves as the lower electrode. A confining ring of some form is then placed on top of the sample into which a pool of mercury is poured which serves as the upper electrode.

When transparent samples are floated in this way it is observed that if a sample is simply laid flat upon the surface of the mercury, con-

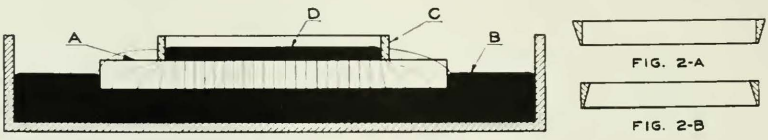


Fig. 2

siderable air is trapped between the sample and the mercury. However, if the sample is lowered obliquely on to the surface of the mercury the air can usually be eliminated. When the mercury has been poured on top of the sample it is impossible to see whether the air has been eliminated from the upper surface of the sample or not. This is sometimes considered a serious objection to the use of this form of electrode especially when used with very thin samples. It may also be questioned whether or not the mercury completely fills the angle between the sample and the inside surface of the ring.

Another point to be considered in the use of this type of electrode is whether the confining ring should be of conducting or of insulating material. The ideal condition, of course, would be to have the confining ring eliminated or, what is the same thing, to have it infinitely thin. Practically, however, something is required to confine the mercury to a definite area. Consider for a moment the arrangement shown in Fig. 2. *A* is the sample under test, *B* is the mercury

constituting the lower electrode, C is the confining ring and D the mercury of the upper electrode. Consider first the ideal case where the ring is of finite height but infinitely thin. This corresponds to the condition illustrated in Fig. 1 except that the fringing is increased due to the size of the lower electrode. Suppose now that the ring is made of insulating material and of appreciable thickness. Some of the lines of force then pass through the material of the ring and the results become dependent on the nature and amount of this material. On the other hand, suppose the ring is made of metal and of appreciable thickness. This, of course, increases the effective area of the electrode but this is easily remedied by making the outside diameter of the ring equal to the size of electrode desired. However, that part of the surface which is actually covered by the ring is subject to the same error as in the case of plain metal electrodes, namely, that if the surface of the sample is not a true plane there will be a gap between the ring and the sample. One remedy for this is to make the ring thin enough so that the area affected is negligible compared with the total. This may not seem to be consistent with suitable rigidity. However, if a ring of at least 4½ inches diameter is used it can be as much as 1/16 inch thick and therefore quite rigid without serious error. In this case the area of the ring is about 5.6% of the total area and assuming a 10% error due to the air-gap as in the case of the plain electrodes the result would be a net error of .56% in the dielectric constant and a correspondingly slight error in the power factor.

Another means of reducing this error without sacrificing the rigidity of the ring is to bevel the edge of the ring either on the outside or inside as shown in Figs. 2-A and 2-B, respectively. In this way the area covered by the ring can be reduced to less than 1% of the total area and the air-gap error to less than 0.1%. However, if the ring is beveled on the outside the outer surface of the ring is no longer perpendicular to the surface of the sample and a slight increase in the stray field from this surface is produced. On the other hand, when the ring is beveled on the inside, the angle between the inner surface of the ring and the sample is less than a right angle and any tendency of the mercury not to fill this angle will be increased. However, both of these factors are probably but little affected by the slight bevel which is sufficient to reduce the thickness of the edge to a small value.

Another form of mercury electrode designed especially to eliminate all errors due to air bubbles has been used by Dye and Hartshorn for measuring the dielectric constant of mica in very thin sheets.[3]

[3] Proceedings of the Physical Society of London, December 15, 1924.

In this arrangement the sample is clamped in a vertical position between two ebonite plates. These plates are recessed in such a way that a closed cell is formed on each side of the sample. An air vent is provided at the top of the cell and the mercury is introduced through a capillary U tube connected to the bottom of each cell. Thus the mercury is caused to rise slowly along the surface of the sample displacing the air completely. Also a slight head of mercury can be maintained in order to force it into all angles. However, this electrode is subject to the errors of the confining ring of insulating material in an exaggerated form on account of the mercury electrodes being entirely enclosed by the ebonite clamping plates. This would unquestionably lead to appreciable errors in the power factor especially in the case of thick samples of low loss material.

Foil Electrodes

Another form of electrode which has been widely used consists simply of a sheet of tinfoil applied to either side of the sample usually with a thin film of wax or petrolatum to serve as an adhesive. This has the advantage that the thickness of the electrode can be made negligible thereby practically eliminating the error due to the field from the vertical surface of the electrode passing partly through air and partly through the sample. The capacitance from the upper surface of the electrode to ground however is not eliminated.

The question of the size of the electrodes also arises. If both the electrodes are extended entirely to the edge of the sample the edge correction for the capacitance is greatly reduced since the fringing all takes place in air having a dielectric constant of 1 instead of the higher dielectric constant of the sample. On the other hand the fringing through the air does produce a small effect on the power factor which does not exist when the fringing is all through the sample. However, the biggest objection to this arrangement from the practical standpoint probably is that the samples very frequently are not uniform in thickness near the edge and this makes it difficult to determine the effective thickness of the sample.

Another possibility is to make both electrodes the same size but smaller than the sample. This should result in a comparatively small edge correction but requires careful manipulation to insure the electrodes being exactly opposite each other. The simplest arrangement from a convenience standpoint is to have one large and one smaller electrode. This however, results in a further increase in the edge correction.

Conducting Paint Electrodes

Another form of electrode which might be considered as a variation of the foil electrode consists of a coating of conducting paint on either side of the sample. In general the conditions are the same as with foil electrodes. A possible advantage might be more intimate contact with the sample and a possible disadvantage is that the film may have sufficient resistance to materially affect the power factor measurement. Many metallic paints are almost entirely non-conducting and are therefore entirely unsuitable for this purpose. If suitable conductivity is obtained, the discussion of foil electrodes given above may be applied to this type also.

Fixed Gap Electrodes

Another type of electrode occasionally referred to [4] consists essentially of a parallel plate air condenser the capacity of which is measured first alone, and then with the sample of insulating material inserted in the air-gap but not necessarily filling the gap completely. From these measurements and the known dimensions of the air condenser and sample the dielectric constant and power factor of the sample can be computed. This arrangement is capable of high accuracy if the dimensions of the sample and the thickness of the gap are known with sufficient accuracy. The computations are not as simple as for the other electrodes referred to and a slight error in determining the thickness of the sample or gap results in a much larger error in the final results. Therefore, this method does not seem to offer any marked advantages over the simpler forms.

Evaluation of Errors

Having now discussed in a general way the various types of errors and their probable effect in connection with various types of electrodes, an attempt will be made to determine by experiment the magnitude of the more important of these errors with respect to certain definite types of electrodes. From the discussion already given, it appears that some form of the mercury or foil electrode should be the most suitable for general use. Hence this investigation will be confined to these two general types.

The first question, namely, that of reproducibility, is one which cannot be determined by a single observer except as it applies to his own particular method of manipulation. Since the results obtained

[4] A. Campbell, Proceedings of the Royal Society, Volume 78, Page 196 and Dye and Hartshorn Local Citation.

may depend considerably on the skill and patience exercised in the handling of the samples and electrodes, they may vary considerably with different observers. Hence, a comprehensive discussion of this point is beyond the scope of this paper. It has been the experience of the writer, however, that there is little choice between the two in this respect and that a decision between them rests primarily on other factors.

Since the magnitude of the ground capacitances and fringing effects both for foil and for mercury confined by a metal ring can be determined from a single series of tests, such an experiment will now be described.

Samples of insulating material 6 inches square were entirely coated on both sides with tinfoil using petrolatum as an adhesive. After the foils were in place, a 4½ inch circle was described on each foil from the center of the square and cut through so that the inner and outer portions were not in electrical contact, although the separation between them was very small. This left two 4½ inch disc electrodes L and M, on opposite sides of the sample surrounded by the annular pieces N and O respectively. When the inner and outer sections are connected, we have the condition of foil electrodes covering the entire surface of the sample. Then if N is used as a guard ring, it is possible to obtain measurements between L and M under uniform field conditions with all fringing effect eliminated. If N is removed and M and O are connected, we have the condition of one large and one smaller electrode. If a metal ring 4½ inches outside diameter is placed on the foil, we have a condition similar to that of a mercury electrode confined by a metal ring. If both N and O are removed, we have the case of two equal foil electrodes smaller than the sample. All of these variations can be obtained without any variation whatever in the conditions of contact between L and M and the sample and therefore are directly comparable.

The method of making these measurements [5] by means of a completely shielded capacitance and conductance bridge [6] is the same as that described by Campbell for the measurement of direct capacitance and will not be described here. The measurements were made at a frequency of 1,000 cycles as fewer difficulties are encountered than at radio frequencies and the general results are the same for any frequency. The capacitances were balanced to 0.1 mmf. or better, and the conductances to 0.0001 micro-mho.

[5] G. A. Campbell, *Bell System Technical Journal*, July, 1922 and *Journal of the Optical Society of America and Review of Scientific Instruments*, August, 1922.

[6] G. A. Campbell, *Electrical World*, 43, 1904, 647-649.

The complete series of measurements made on the samples discussed above was as follows:

1. Grounded capacitance of L+N to M+O. This includes the stray capacitance of the upper electrode to ground and also any fringing effect in the air around the edges of the sample.
2. Direct capacitance of L+N to M+O. This eliminates the stray capacitance to ground but includes the above fringing, therefore, 2 minus 1 gives the stray capacitance of the 6″ square electrode to ground.
3. Direct capacitance of L+N to ground using M+O as a shield. This should check 2 minus 1 above.
4. Direct capacitance from L to M using N as a guard ring and eliminating all fringing and ground capacitance.
5. Direct capacitance from L to M+O with N removed. This includes the fringe effect from a small to a large electrode but eliminates the capacitance of L to ground. Hence, 5 minus 4 gives the fringe effect.
6. Grounded capacitance of L to M+O with N removed. This includes both fringe effect and capacitance of L to ground. Hence, 6 minus 5 gives the capacitance of L to ground.
7. Same as 6 with 4½″ metal ring on top of foil L. 7 minus 6 gives the added capacitance due to the ring.
8. Direct capacitance of L to ground using M+O as shield. This should check 6 minus 5 above.
9. Direct capacitance of L to M with N and O removed. This includes the fringe effect between equal electrodes but eliminates the capacitance of L to ground.
10. Grounded capacitance of L to M with N and O removed. This includes the fringe effect and ground capacitance for equal electrodes. 10 minus 9 equals capacitance of L to ground with equal electrodes.
11. Direct capacitance of L to ground using M as shield. This should check 10 minus 9 above.

The results of these measurements on a number of samples including several thicknesses of phenol fibre, hard rubber and glass are given in Table I. In all cases, the capacitance of the leads was measured separately and deducted. The differences between readings as indicated above, and the corresponding check readings are tabulated in Table II.

In using the results tabulated in Table II the directly measured values of stray capacitance to ground (items 2, 6, 12) should be

considered much more reliable than those determined by differences (items 1, 5, 11). The degree of agreement between the two should be considered more as a check on the accuracy of the individual values from which the latter are derived than as a check on the former.

Hence, we may consider 3.3 mmf. (item 2, Table II) as the stray capacitance for the 6 inch square and about 2.3 mmf. (item 12) for the 4½ inch circle. This checks fairly well with the theoretical value of $\frac{R}{\pi}$ (CGS units). If the square is considered equivalent to a circle of equal area the value of $\frac{R}{\pi}$ is equivalent to about 3.03 mmf. For the 4½ inch circle the value of $\frac{R}{\pi}$ is 2.0 mmf. The measured values should be somewhat higher than the theoretical since the shielded bridge and other apparatus comprise a considerable mass of grounded metal at no great distance from the sample.

The fringe effect for the equal circular electrodes may be compared with values computed from Kirchhoff's formula

$$C=\frac{r}{4\pi}\left(\log_e\frac{16\pi(b+t)r}{b^2}+\frac{t}{b}\log_e\frac{b+t}{t}-3\right),$$

where C is the fringe effect and r and t are the radius and thickness respectively of the electrodes and b is their separation, all in CGS units. For very thin electrodes this reduces to

$$C=\frac{r}{4\pi}\left(\log_e\frac{16\pi r}{b}-3\right).$$

Values for this expression reduced to a percentage correction are listed as item 13 in Table II and are plotted in Fig. 3. It will be noted that the observed effect is at least a third less than that computed. It should be borne in mind, however, that Kirchhoff's formula applies primarily to electrodes in air. To completely simulate this condition with a solid dielectric would require that the electrodes be completely surrounded by a considerable thickness of the dielectric. If the difference noted above is due to lines of force which pass partly through air and partly through the sample, this difference should be greatest for a sample of high dielectric constant and diminish as the dielectric constant of the sample approaches that of the air. It is seen from Fig. 3 that this is the case, the curve for hard rubber having a dielectric constant of 3 being nearer to the computed curve than that for glass having a dielectric constant of 7.7. The anomalous

shape of the curve for phenol fibre is apparently due to a non-uniformity in the samples which will be discussed later.

That the above results are materially affected by flux passing partly through air and partly through the sample was proved by an additional test. The $\frac{3}{8}$" phenol fibre sample with $4\frac{1}{2}$" discs on each side

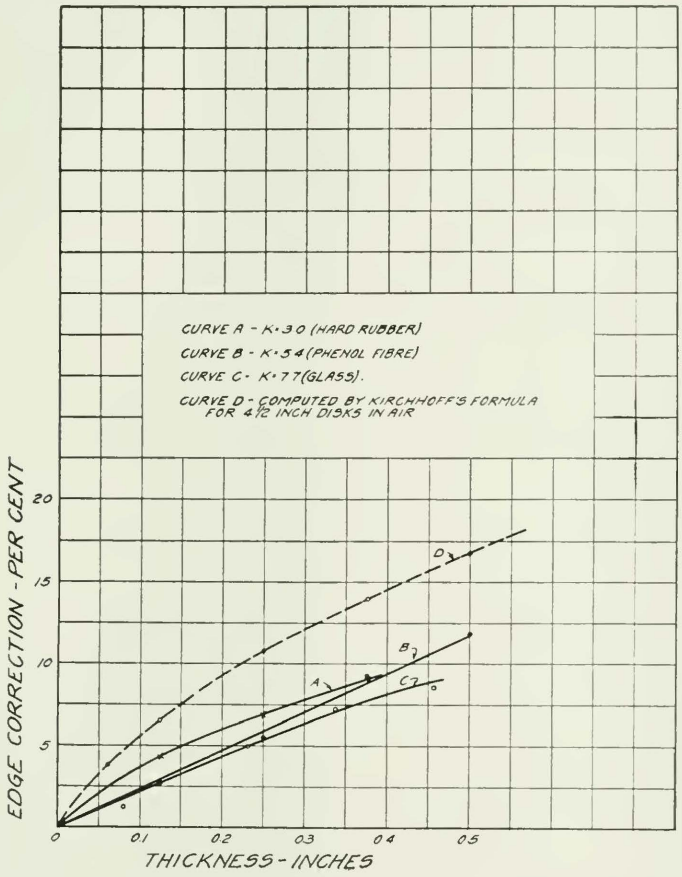


Fig. 3—Percentage edge correction for samples 6 inches square with foil electrodes both $4\frac{1}{2}$ inches in diameter

was first measured in air. Since the upper electrode could not be entirely covered and still obtain contact with it, the $\frac{1}{4}$" phenol fibre plate was laid on top of it so as to cover just one-half of the sample and electrode. This caused an increase of 1.4 mmf. Using the $\frac{1}{4}$" plate in the same way to cover one-half of the lower surface of the sample and electrode, the increase was .7 mmf. showing that

when one electrode is grounded the effect is not symmetrical. Covering both sides of the sample should, therefore, cause an increase of 4.2 mmf. or 8.6% of the capacitance of the sample. Adding this to the 9.3% fringe effect already determined makes a total of 17.9% as compared with 13.9% by Kirchhoff's formula. At least part of this difference is also due to the non-uniformity referred to above which is very marked in this sample.

A similar test was made on the $\frac{1}{8}''$ phenol fibre sample. However, this sample was somewhat warped so that there was an appreciable air-gap between the electrode and the cover plate. The total effect computed as above was 2.6%, which added to the 2.8% already determined makes a total of 5.4% or slightly under the value computed by Kirchhoff's formula. This, no doubt, is accounted for by the air-gap.

The agreement with Kirchhoff's formula is reasonably good, therefore, for disc electrodes completely surrounded by dielectric, but it is evident that the formula does not apply to the case of disc electrodes applied to sheet materials.

The fringe effect for the $4\frac{1}{2}''$ upper circle and 6″ lower square electrodes is shown in Fig. 4. It is found to be about $2\frac{1}{2}$ times as large as for the equal $4\frac{1}{2}''$ electrodes. It also varies somewhat with the dielectric constant of the sample, being greater for the lower dielectric constant. The anomalous behavior of the phenol fibre samples is shown in this figure also.

From Item 8 of Table II, it is seen that when a shallow metal ring is placed on the upper disk to simulate the conditions of a mercury electrode, a further increase of from 2 to 4% of the true capacitance of the sample takes place. This likewise is greater the thicker the sample and the lower its dielectric constant. A similar test for the increase in capacitance due to vertical height of the metal ring was made for the more exaggerated case of a 4″ disk and a ring $\frac{3}{4}''$ high; the lower electrode being 6″ square. In this case the increase varies from $2\frac{1}{2}$% for $\frac{1}{8}''$ phenol fibre to 8% for $\frac{3}{8}''$ hard rubber. This shows the importance of keeping the vertical dimension of the metal ring as small as possible.

Insulating Ring

In order to determine the corresponding effect when an insulating ring is used for confining mercury, a somewhat similar test was made. Several different rings were used as follows: ring No. 1 is $\frac{3}{4}''$ high cut from hard rubber tubing having a $\frac{1}{8}''$ wall with the edges cut square. Ring No. 2 was the same as above except that the edge was

beveled on the outside as in Fig. 2-A to a thickness of 1/64″. Ring No. 3 was cut from phenol fibre sheet $\frac{1}{8}$″ thick and had a radial width of $\frac{3}{8}$″. Since rubber tubing of the desired size was not available the rubber rings were somewhat smaller than $4\frac{1}{2}$″ in diameter but in all cases the foil electrode with which they were used was cut

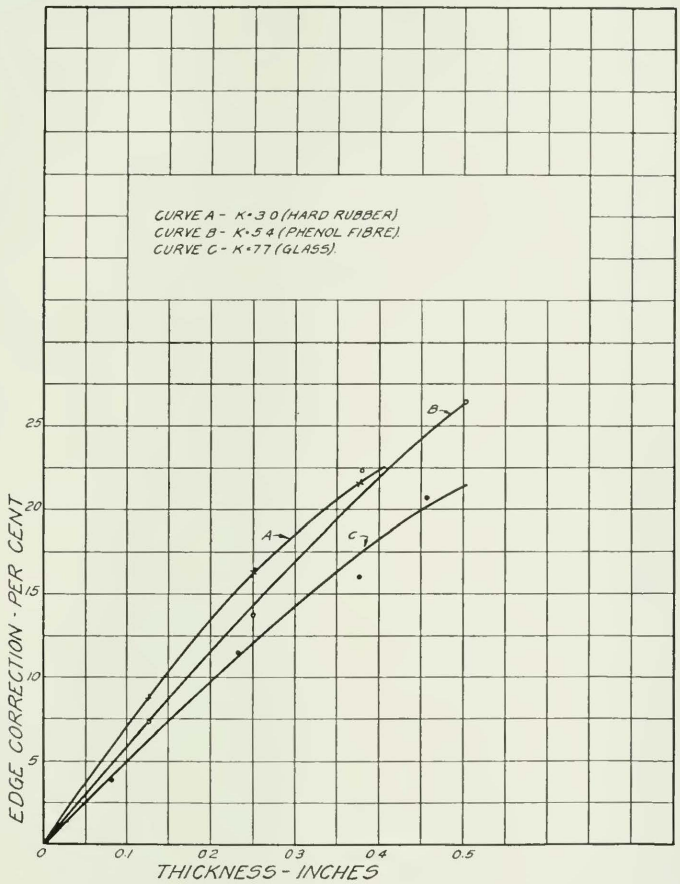


Fig. 4—Percentage edge correction for samples 6 inches square with foil electrodes upper, $4\frac{1}{2}$ inches diameter, lower, 6 inches square

to exactly the same size as the inside diameter of the ring and the results were reduced to the equivalent value for $4\frac{1}{2}$″ diameter.

In this test in order to simulate the depth of the mercury without changing the electrode contacts, the lower part of the inner surface of the ring was coated with foil to a height equal to the assumed depth of the mercury. This was taken as $\frac{1}{8}$″ in each case which is

about the minimum depth of mercury that can be depended upon to cover the surface of the sample and fill the angle between the sample and the ring. This, of course, has slightly less effect than if the entire surface of the electrode were raised as is the case when mercury is poured into the ring, but the difference is probably negligible. The procedure was the same as previously described for the metal ring. The lower foil electrode covered the entire surface of the sample. The upper foil is the same diameter as the inside of the insulating ring. (In the case of the metal ring the foil was the same size as the outside diameter of the ring at the lower edge). The capacitance is measured first between the foils alone and then with the ring superimposed on the upper foil. The change should represent very closely the increased edge effect due to a depth of $\frac{1}{8}''$ of mercury in the ring.

The results of this test are tabulated in Table III. It will be noticed that the change in capacitance due to the beveled rubber ring is about the same as for the metal ring, while the change due to the square edged ring is materially greater than for the metal ring. The flat phenol fibre ring produces a change in capacitance two or three times as great as the metal ring and the apparent power factor for the rubber sample is more than doubled. This, of course, is due to the dielectric loss in the ring itself and therefore the lower the power factor of the sample under test the greater is the proportional error. While the rubber rings had no appreciable effect on the power factor of the samples tested, it is believed that in the case of very low loss materials such as fused quartz the effect might still be appreciable. Since the insulating rings under the best conditions are no better than a metal ring and under poor conditions are very much inferior, it is believed that in general metal rings will be found more satisfactory for confining mercury electrodes.

Thus far we have considered the experimental data primarily with regard to the capacitance and dielectric constant. Table IV shows the power factors computed from the conductance readings corresponding to readings 1, 4, 6 and 10 in Table I. These values were computed from the capacitance and conductance values for the various types of electrodes and without correction for edge effects or ground capacitances. The values for hard rubber illustrate fairly well the variation for different electrodes which would be expected on the basis of the preliminary discussion. The value obtained with the $4\frac{1}{2}''$ circle and guard ring should be the true value. The other values should be slightly lower on account of the additional capacitance without corresponding power less due to the flux which passes partly and entirely through air. In general these variations are small

and almost beyond the limit of accuracy of the measurements. However, a careful analysis of the results (omitting the $\frac{3}{8}''$ phenol fibre sample which will be discussed separately) seems to indicate that the power factor values obtained with the two $4\frac{1}{2}''$ circles agree slightly better with those obtained with the guard ring electrode than do those obtained with any of the other electrodes.

In the case of the $\frac{3}{8}''$ phenol fibre sample the variations with different electrodes are much greater than the probable inaccuracy of the measurements. Apparently, they can only be attributed to non-uniformity of the material in different parts of the sample. Therefore, a special test to determine this fact was made on this sample. By interchanging the connections to the guard ring and the $4\frac{1}{2}''$ center electrode, it is possible to measure the capacitance and conductance of the outer part of the sample without including the center part. The sum of the values of the two parts checks well with the value for the entire 6″ square. These results show that while the inner part of the sample has a power factor of 1.97%, the power factor of the outer part is 3.17% or approximately 60% higher. The corresponding dielectric constants are 5.08 and 5.48, respectively. The reason for this difference probably is that since the material is of a laminated nature moisture penetrates more readily from the edges than from the sides of the sample and thus causes a progressive variation of the electrical characteristics from the edges to the center of the sample. It is obvious that when the two $4\frac{1}{2}''$ electrodes are used without guard rings, some of the outer part of the sample is included due to the fringe effect and that when the $4\frac{1}{2}''$ upper and 6″ lower electrodes are used still more of the outer part is included for the same reason, and the values obtained are increased accordingly.

As previously mentioned, it is probable that this non-uniformity is responsible for the different shape of the edge effect curve for phenol fibre as compared with those of hard rubber and glass and it is almost certainly the cause of the point representing this particular sample being exceptionally high. Since there are wide variations in the power factors of the different samples of both phenol fibre and glass it is possible that there are minor variations through the sample due to causes other than moisture and that these may account for some of the other apparent irregularities in the results.

Method of Applying Corrections

While percentage values are the most convenient for discussion of the relative importance of the various corrections involved in the use of a given type of electrode, the absolute values of these corrections

in micro-microfarads are probably somewhat more convenient for actual use in making the necessary calculations.

Consider the case of mercury electrodes with the lower electrode grounded. On the basis of the foregoing discussion the total capacitance C which is measured may be considered as made up of four parts, namely, C_x the capacitance between the electrodes which would exist under uniform field conditions as when a guard ring is used. C_{e1} the edge effect which would exist if the upper electrode were a thin disk. C_{e2} the additional edge effect due to the height of the metal ring. C_g the capacitance to ground of the upper surface of the electrode. The dielectric constant $K=4.46\,\frac{C_x d}{A}$ where C_x is in micro-microfarads d and A are the thickness and area of the sample in inches and square inches, respectively. To compute K, C_x must therefore be obtained from the relation

$$C_x = C - (C_{e1} + C_{e2} + C_g).$$

Values for C_{e1}, C_{e2}, C_g are given in the tables for certain particular cases and may be determined in a similar manner for any other cases. For foil electrodes, conditions are similar except that C_{e2} is zero. C_{e1} which is the largest of the corrections is plotted in Fig. 5 for various values of K and several thicknesses of the sample as taken from the previous tables. It will be seen that in order to apply this correction the approximate value of the dielectric constant of the sample must be known in advance. This can always be obtained by making a preliminary computation neglecting the corrections entirely.

As an example of the above method suppose measurements have been made on a $\frac{1}{8}''$ sample of material having a dielectric constant of about 4.5 using mercury electrodes with a shallow metal ring. From Curve A, Fig. 5 we get 9.2 mmf. for C_{e1}. C_{e2} is estimated by interpolation from item 7 of Table II at about 2.7 mmf. C_g is taken as the average for item 6 of Table II or 0.8 mmf. This makes a total of 12.7 mmf. to be subtracted from the measured capacitance in order to obtain C_x from which the dielectric constant is computed. If $4\frac{1}{2}''$ foil electrodes were used C_{e1} would be taken from Curve B, Fig. 5 as 4.0 mmf. and C_g from the average of item 11, Table II as 2.4 mmf. making a total correction of 6.4 mmf.

The values given in Fig. 5 for C_{e1} are, of course, applicable only to a given size of electrode, namely $4\frac{1}{2}''$ in diameter. If the edge effect capacitance is considered equivalent to that of an additional ring electrode surrounding the main electrode, this capacitance would be proportional to the mean radius of this ring, or to the radius of the

upper electrode plus one-half of the width of the ring. If this equivalent ring is of constant width for various sizes of electrodes, the edge correction would be proportional to the radius of the electrode plus a constant. As Fig. 5 shows that the correction in micro-microfarads does not vary greatly with the thickness of the sample, the width of

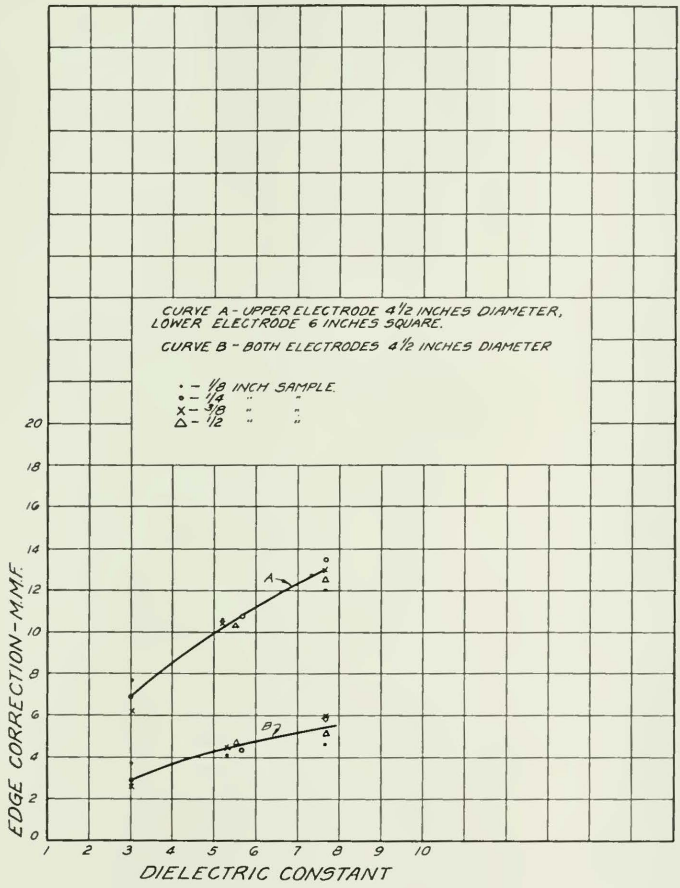


Fig. 5—Edge correction versus dielectric constant for samples 6 inches square with foil electrodes

the above hypothetical ring must be approximately proportional to the thickness of the sample. Computations for several samples show that the width of this ring is of the order of twice the thickness of the sample. Hence it appears that as a first approximation the edge correction for various sizes of electrodes may be taken as pro-

portional to $r+t$, r being the radius of the electrode and t the thickness of the sample. Table V shows a series of measurements on one sample with electrodes of several different sizes. In this table the edge correction C has been taken as the difference between the capacitances measured with and without a guard ring. This includes the direct capacitance of the upper electrode to ground. It will be seen that while the value of $\frac{C}{r+t}$ is not exactly constant it varies much less than $\frac{C}{r}$. Hence if C_1 is the edge correction for an electrode of radius r_1 the edge correction for an electrode of radius r_2 under similar conditions would be approximately

$$C_2 = C_1 \frac{r_2+t}{r_1+t}.$$

This, of course, applies only to the case of one large and one smaller electrode. For the case of the equal electrodes t in the above expression probably should be multiplied by a constant of the order of 0.4.

Tables for the Article

By E. T. Hoch

TABLE 1

Capacitance Readings as Described under "Evaluation of Errors" Micromicrofarads

Reading No.	Phenol Fibre				Hard Rubber			Glass			
	½ in.	⅜ in.	¼ in.	⅛ in.	⅜ in.	¼ in.	⅛ in.	.455 in.	.338 in.	.232 in.	.080 in.
1	94.7	116.3	*	335.0	69.5	100.5	200.4	147.0	190.2	278.0	790.8
2	*	113.1	*	332.2	66.7	97.8	195.1	*	*	*	*
3	*	3.2	*	3.3	*	3.3	3.3	*	*	*	*
4	39.7	48.6	60.2	146.6	28.9	42.3	86.4	61.4	82.0	118.4	344.6
5	(50.1)	59.1	91.0	157.2	35.1	49.1	94.1	(74.0)	(95.0)	(131.9)	(356.6)
6	50.8	59.6	92.4	157.7	35.7	49.4	94.4	74.7	95.7	132.6	357.3
7	*	60.9	*	160.8	36.9	51.3	96.5	*	*	*	*
8	*	0.8	1.3	1.0	0.7	0.6	0.7	*	*	*	*
9	(44.7)	53.1	84.6	150.7	31.5	45.2	90.1	(66.6)	(88.0)	(124.3)	(349.2)
10	46.7	55.7	87.9	153.1	33.6	47.2	91.2	66.9	90.3	126.6	351.5
11	*	2.6	3.0	2.3	2.3	2.3	2.1	*	*	*	*

* Not measured.

NOTE: Readings 5 and 9 given in parentheses are derived from readings 6 and 10 respectively using average values for readings 8 and 11.

TABLE 2

Values Computed from Table 1 and Comparison with Measured Values

Value Computed (mmf. unless otherwise stated)	Phenol Fibre				Hard Rubber			Glass			
	½ in.	⅜ in.	¼ in.	⅛ in.	⅜ in.	¼ in.	⅛ in.	.455 in.	.338 in.	.232 in.	.080 in.
(1) Cap. of 6 in. square electrode to ground—Reading 1–2 (Table 1)	*	3.2	*	2.8	2.8	2.7	5.3				
(2) Ditto by direct meas. Reading 3	*	3.2	*	3.3	*	3.3	3.3				
(3) Fringe effect 4½ in. circle to 6 in. square—Reading 5–4	10.4	10.5	10.8	10.6	6.2	6.8	7.7	12.6	13.0	13.5	12.0
(4) Ditto in per cent	26.2	22.4	13.5	7.2	21.4	16.1	8.8	20.5	15.8	11.4	3.4
(5) Stray cap. to grd. from 4½ in. upper with 6 in. lower electrode (6–5)	*	0.5	1.4	0.5	0.6	0.3	0.3				
(6) Ditto by direct meas. Reading 8	*	0.8	1.3	1.0	0.7	0.6	0.7				
(7) Additional capacitance due to metal ring Reading 7–6	*	1.3	*	3.1	1.2	1.9	2.1				
(8) Ditto in per cent	*	2.7	*	2.1	4.1	2.6	2.4				
(9) Fringe effect 4½ in. circle to 4½ in. circle—Reading 9–4	4.7	4.5	4.4	4.1	2.6	2.9	3.7	5.2	6.0	5.9	4.6
(10) Ditto in per cent	11.8	9.3	5.5	2.8	9.0	6.8	4.3	8.4	7.3	5.0	1.3
(11) Stray cap. to grd. from 4½ in. upper with 4½ in. lower electrode (10–9)	*	2.6	3.3	2.4	2.1	2.0	1.1				
(12) Ditto by direct meas. Reading 11	*	2.6	3.0	2.3	2.3	2.3	2.1				
(13) Fringe effect by Kirchhoff's formula for 4½ in. circles in air (per cent)					13.9	10.7	6.6				

*Not determined.

TABLE 3

Effect of Various Types of Rings for Confining Mercury Electrode

	Increase in Capacitance Per Cent		Power Factor Per Cent	
	1/8 in. Glass	3/8 in. Hard Rubber	1/8 in. Glass	3/8 in. Hard Rubber
Without ring	0.0	0.0	2.14	.43
With metal ring 3/16 in. high, outside bevel	2.1	4.1	2.11	.45
With square edged hard rubber ring	2.7	7.0	2.11	.45
With bevel edged hard rubber ring	1.8	4.9	2.13	.44
With flat phenol fibre ring	4.7	12.6	2.45	1.08

TABLE 4

Power Factors as Measured with Different Electrodes

Electrodes	Phenol Fibre				Hard Rubber			Glass			
	1/2 in.	3/8 in.	1/4 in.	1/8 in.	3/8 in.	1/4 in.	1/8 in.	.455 in.	.338 in.	.232 in.	.080 in.
6 in. squares	2.40	2.24	1.95	2.18	0.43	0.41	0.45	1.67	2.46	2.18	2.94
4½ in. circle with guard ring	2.56	1.96	2.05	2.19	0.44	0.47	0.44	1.81	2.43	2.21	2.86
4½ in. upper, 6 in. lower	2.50	2.16	2.08	2.20	0.40	0.39	0.46	2.00	2.43	2.14	2.85
Ditto with metal ring	*	2.12	*	2.18	0.39	0.39	0.46	*	*	*	*
4½ in. upper and lower	2.58	2.14	2.11	2.20	0.44	0.45	0.44	1.80	2.50	2.20	2.90

* Not measured.

TABLE 5

Measurements on 1/4 In. Phenol Fibre with Electrodes of Various Sizes

Radius of Upper Electrode In.	Cap. With Guard Ring Mmf.	Cap. Without Guard Ring Mmf.	Edge Correction C Mmf.	C/r	C/r+t	Diel. Const. Computed from 2nd Col.
.51	4.25	7.8	3.55	7.0	4.7	5.7
1.00	15.9	21.9	6.0	6.0	4.8	5.58
1.50	35.4	44.1	8.7	5.8	5.0	5.50
2.00	62.6	73.2	10.4	5.2	4.6	5.50
2.26	82.1	92.9	10.8	4.8	4.3	5.60

NOTE: Sample was 6 in. square with lower electrode covering entire lower surface.

Load Carrying Capacity of Amplifiers

By F. C. WILLIS and L. E. MELHUISH

Synopsis: This paper describes the adaptation of the cathode ray oscillograph to the determination of the overload point of vacuum tube amplifiers. Using the input voltage to produce a horizontal deflection, and the output voltage or current to produce a vertical deflection, the amplifier performance is readily determined by noting the resulting figure on the fluorescent screen. So long as the figure is virtually a straight line or an obviously undistorted ellipse, it was found that the amplifier output is free from harmonics. As soon as overloading begins, the oscillogram shows either a sharp bend at either or both extremities of the line or apparent distortion of the ellipse. The method has the advantage of being quick.

IN any device used for the amplification of a complex electrical wave such as that necessary for the transmission of speech or music, distortion may arise in two ways. (a) The amplification may not be the same for all frequencies in the band to be transmitted. (b) The relationship between input voltage and output current may not be such as can be described by a straight line when r.m.s. values are plotted against each other.

Considering the distortion due to cause (b) alone it is true that for most practical devices the relationship between input and output can be described by a curve which is approximately straight for a portion of its length but as the amplitude of the wave to be transmitted increases, operation is over a longer portion of the curve and it ceases to be possible to regard the characteristic curve as straight. It therefore becomes necessary to determine for any design the maximum energy that the system can carry without noticeable distortion from this cause. For a system intended to transmit speech or music the final decision as to how much distortion is permissible must depend upon the judgment of the expert listener but analytical methods are of service in establishing reference points by measurements which can be duplicated without reference to any particular person. The purpose of this paper is to describe some work undertaken for this purpose.

Departure from the ideal straight line relationship between input and output results in the production of harmonic overtones of all the frequencies present in the input and in the production of beat notes between frequencies if there is more than a single frequency in the input. It follows from this that if a complex wave is passed through a device having a characteristic of this nature the output will differ from the input in the proportion of the different frequencies and may contain frequencies that were not present in the input at all. A change in the proportion of the different frequencies may be partly

due to cause (a) but the introduction of new frequencies can only be due to cause (b). As is now known (BELL SYSTEM TECHNICAL JOURNAL October, 1923) the response of the ear itself is non-linear so that subjective harmonics and sum and difference tones are heard by every listener. Under good conditions, therefore, the distortion produced by the non-linear transmission of the amplifier will be so small compared to that produced in the ear itself that it will not be noticed. On the other hand, it may be so great as to render unrecognizable the speech or music being transmitted. This condition will be familiar to everyone who has been compelled by an enthusiastic friend to listen to a heavily overloaded radio receiver.

The principal parts of a vacuum tube amplifier where one might expect to find the non-linear response under consideration are in the magnetic circuits of the transformers and retard coils and in the vacuum tubes. Generally, in the amplifiers considered in this paper, the design of the transformers is such that the magnetic flux density is small so that the magnetization curve is practically a straight line and very little distortion is to be expected from this cause. On the other hand the $E_c - I_b$ characteristic of the vacuum tube is approximately straight for only a small portion of its length and has a pronounced curvature in the usual working range. This is the case even for well designed circuits where under proper operation there is no possibility of the grid of the tube drawing current. It is therefore, in the characteristics of the vacuum tubes that the principal source of trouble of this nature is to be looked for. That this anticipation is justified will be shown by the results described in this paper.

The relationship between output and input of a vacuum tube has been studied from a mathematical viewpoint and formulae have been established by which the resultant output for a given input may be calculated provided the tube parameters and circuit impedances are precisely known. For any commercial amplifier having several stages the measurement of these quantities and the necessary calculations would be a slow procedure. By experimental methods it is possible to determine directly and quantitatively the distortion that occurs in any particular case. This has been done for a number of amplifiers under various load conditions with a view to establishing convenient criteria by which it is possible to determine quickly and easily how much energy any amplifier will transmit without serious distortion. The amplifiers dealt with were all audio-frequency amplifiers so that no questions of radio frequency amplification or of intentional rectification or modulation are considered and the measurements were all made with single frequency inputs as this naturally forms the basis

for a more complete analysis of the problem. Three kinds of measurements were made.

1. The gain of the amplifier was measured under load conditions which varied from a point well below its carrying capacity to a point where it was obviously overloaded. The results from tests of this kind show that the gain is uniform at low outputs, begins to fall off when the output reaches a certain level and falls off more and more rapidly as the load is further increased. The point at which the gain begins to fall off has sometimes been taken as a criterion of the load carrying capacity of the amplifier.
2. With a single frequency (1,000 c.p.s.) input to the amplifier the output was analyzed at a number of points along the load-gain curve and the percentage of harmonic to fundamental in the output plotted against the same scale of energy output as for the load-gain curve.
3. The input voltage was made to produce a horizontal deflection in a cathode ray oscillograph while the output voltage or current was made to produce a vertical deflection. In effect this is a convenient means of drawing the input-output characteristic of the amplifier so that its general curvature and the loads at which any sudden changes of curvature occur may be easily observed. For an amplifier that produced no distortion or phase shift, the resultant figure would be a straight line whatever the wave shape of the input. If there were phase shift but no distortion the result would be an ellipse or circle depending on the phase and amplitude relationships of the input and output provided the input were a pure frequency. In general for the practical case the result is a distorted ellipse showing that the wave undergoes both distortion and change of phase in passing through the amplifier. With increasing load the distortion becomes more and more pronounced.

The gain measurements were made by methods in principle the same as those embodied in standard gain measuring sets of the Bell System, and while it is not the intention of this paper to give a detailed description it may be desirable to give a brief statement of the principles involved. For the purposes of this paper the gain of an amplifier is defined as the logarithm of the ratio of the power delivered into its load impedance to the power that would be delivered if the amplifier were removed and replaced by the best possible passive network.

Thus

$$N_{TU} = 10 \operatorname{Log}_{10}\left(\frac{W_o}{W_i}\right).$$

In practice an amplifier is almost always measured between impedances that are pure resistances. These impedances are set up by variable resistance networks so designed that when the current in one mesh of the input network is equal to the current in a mesh of the output network the gain of the amplifier may be read from the settings of the dials and switches controlling the networks. An indicating device which may be switched from the input network to the output network indicates when the currents in the two meshes mentioned are equal. Where sufficient energy is available a thermo-

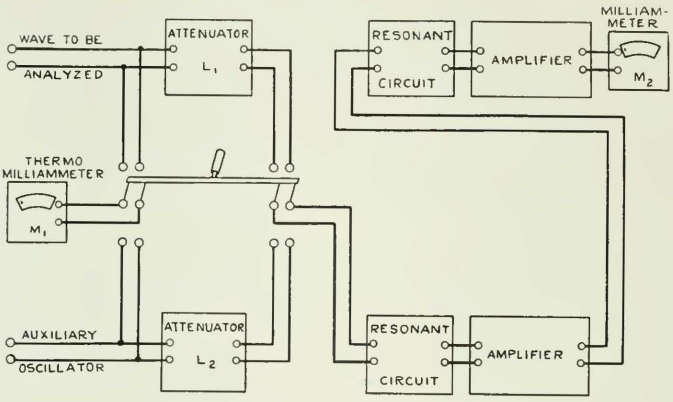


Fig. 1—Circuit for harmonic analysis

couple is used as the indicating device. It follows that measurements are then always in terms of r.m.s. values of the currents employed. So long as these currents are approximately single frequency the r.m.s. value of the whole wave is very nearly equal to the r.m.s. value of the fundamental. In all the experiments described here gain measurements were made by thermocouple and as will be seen later the proportions of harmonics were such that unless the amplifier was overloaded the discrepancy between the r.m.s. measurements made and those that might be made by methods taking account of the fundamental only was usually less than 1%, and in the cases of extreme overload less than 5%.

The harmonic analyses were made by an electrical analyzer whose principles of operation are indicated in Fig. 1.[1] In operating this

[1] The principles of this analyzer are fully described in a paper by Mr. A. G. Landeen to be published in the *B. S. T. J.*

analyzer the resonant circuits of the amplifier detector and the auxiliary oscillator were tuned to the harmonic to be measured. The input currents and the attenuators L_1 and L_2 were adjusted so that the readings of meters M_1 and M_2 did not change when the four-pole, double-throw switch was thrown from one position to the other. It will be seen that the difference between the settings L_1 and L_2 then gives the proportion of harmonic to total r.m.s. value of the output wave of the amplifier. The proportion of harmonic necessary to cause a 1% change in the r.m.s. value of a wave is approximately 14% and for the values obtained in the experiment the r.m.s. value of the output could be taken as equal to the fundamental in the output. The difference between L_1 and L_2 therefore gives the proportion of harmonic to fundamental. In the present work the difference between the frequencies to be separated was large enough to avoid any difficulty in obtaining sufficient resolution by the use of simple resonant circuits.

It is, of course, necessary that for measurements of this kind the current supplied to the amplifier under test should be a single frequency. A vacuum tube oscillator which was known to give a very pure wave was used as the source of current. To obtain sufficient energy for all the measurements made it was necessary to amplify the output current from this oscillator and subsequently filter it to remove the harmonics introduced by the amplifier. Final analysis of the wave applied to the amplifier under test showed in most cases less than 0.2% and in all cases less than 0.5% of third harmonic and less than .1% of all other harmonics. Greater purity could have been obtained at the expense of more time and trouble but this was considered sufficient for the purposes in view. Where necessary a small correction for the harmonic content of the input wave has been applied to the results.

The voltages to operate the cathode ray oscillograph were obtained by a step-up transformer for the input and directly off a resistance potentiometer for the output. The use of a step-up transformer for the output wave is in general undesirable because it introduces phase and amplitude changes which differ for the component frequencies of the wave and thus the transformer itself introduces a distortion which renders the interpretation of the figure as applied to the amplifier distortion more difficult. This limits the method to cases where a minimum of 10 volts is available in the output. For the amplifiers dealt with here this voltage was available and the limitation was not felt. On the input side the step-up transformer has to transmit one frequency only so that the same difficulty does

not occur. Pictures of the figures obtained on the fluorescent screen were taken with an ordinary camera with from one to three-minute exposures. On most of the pictures the horizontal and vertical axes were also recorded on the screen.

The circuits of the amplifiers dealt with and the results of the analyses are graphically presented in the following figures. Amplifier

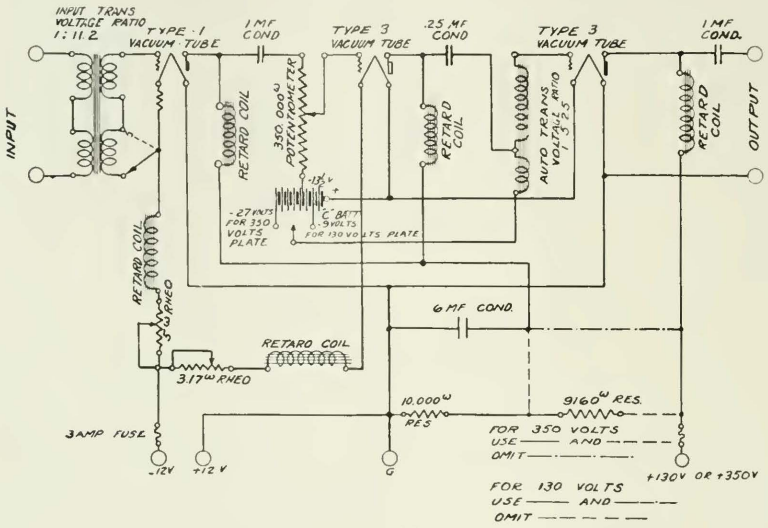


Fig. 2—Amplifier No. 1

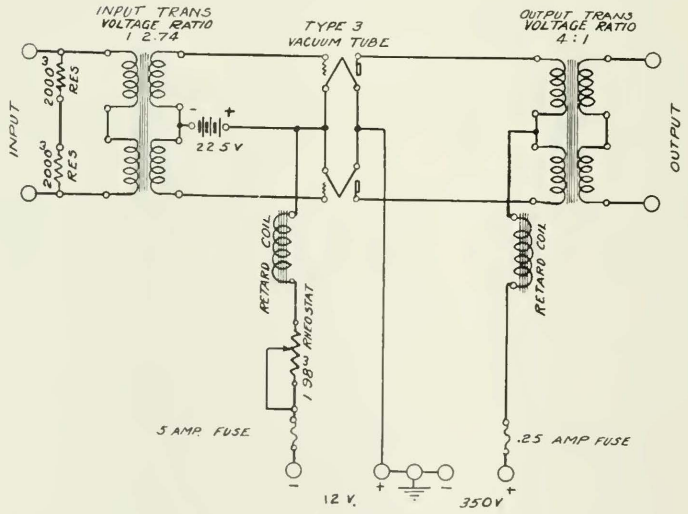


Fig. 3—Amplifier No. 2

No. 1 shown in Fig. 2 is used for amplification from very low up to medium powers in Public Address, Radio Broadcasting and similar systems. Provision is made for operating this amplifier on either 130 or 350 volt anode potential and tests were made under both these conditions. This amplifier is designed to work from an impedance of 200 ohms into one of 4,000 ohms. Amplifier No. 2 shown in Fig. 3

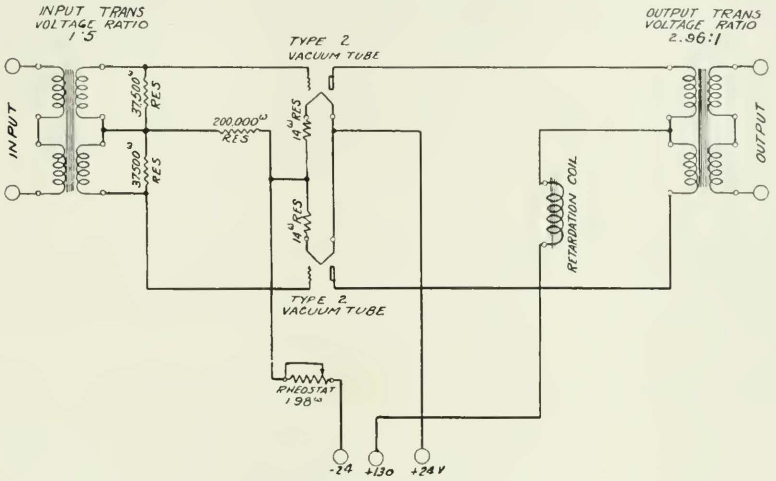


Fig. 4—Amplifier No. 3

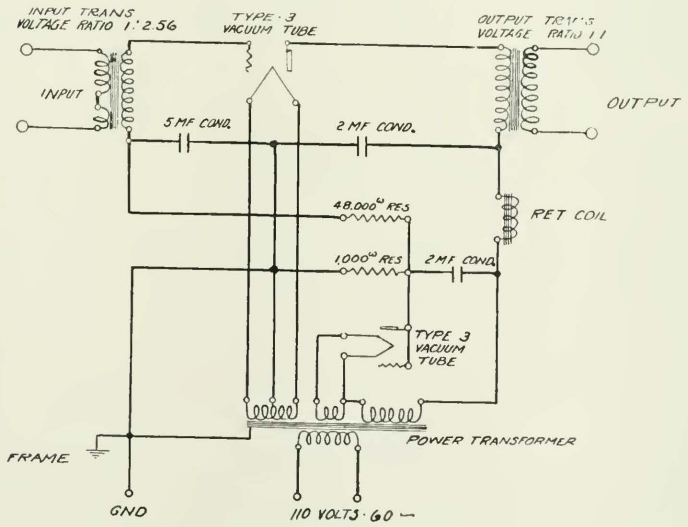


Fig. 5—Amplifier No. 4

follows amplifier No. 1 as the final output stage of a Public Address System and as shown consists of one stage with tubes in the push-pull arrangement. To match amplifier No. 1 its input circuit is designed for 4,000 ohms and its output is stepped down by a transformer to 500 ohms. Amplifier No. 3 shown in Fig. 4, also a push-pull amplifier, is designed as the output stage of an oscillator used in field measurements in the telephone plant. Its input and output impedances are 2,000 ohms and 600 ohms, respectively. Amplifier No. 4 is intended for use in radio reception, amplifying the signals to loud speaker volume and operating from 110 volt, 60 cycle A.C. supply. As shown there is one rectifying tube for converting the 60 cycle a.c. to d.c. for supplying the anode of the amplifying tube. Its rated input and output impedances are 20,000 ohms and 4,000 ohms, respectively.

As will be noted from the diagrams there are employed in these amplifiers three types of vacuum tubes. The normal operating voltages and average characteristics of these tubes are shown in Table I.

TABLE I

	Filament Current	Anode Volts	Grid Biasing Volts	Anode Current Milli-amps.	Amplification Factor μ	Anode-Filament Impedance Ohms.
Vacuum Tube Type 1	1.00	130	− 1.6	0.7	30	60,000
Vacuum Tube Type 2	1.00	130	−20	25.0	2.5	2,000
Vacuum Tube Type 3	1.6	350	−27	25.0	6.5	4,000
Vacuum Tube Type 3	1.6	130	− 9	5.0	6.5	8,000

The harmonic analyses and load-gain curves of the amplifiers under the conditions noted are shown in Figs. 6, 7, 8, 9 and 10. The gain in transmission units and percentage of each harmonic up to the 5th together with the root of the sum of the squares of these percentages being plotted against watts output on a logarithmic scale. The oscillograph pictures taken at various points along these curves are shown in Figs. 11, 12, 13, 14 and 15.

As stated above, the oscillograph figure presents the input-output curve of the amplifier. Furthermore, if there is no distortion in the transformers, the horizontal deflection will be proportional to the alternating grid voltage applied to the first stage while the vertical deflection will be proportional to the alternating component of the plate current in the last stage. The figure drawn is then the dynamic $E_c - I_b$ charactertistic of all the tubes combined. That this is sub-

stantially the case may be seen by inspection of the figures. In the first oscillogram of Fig. 11 there is shown the characteristic for amplifier No. 1 under 130 volts plate supply and .047 watts output where the amplitude of grid voltage is such that no grid becomes positive with respect to the filament nor is the plate current reduced to zero at any part of the cycle. Under these conditions the tube characteristic as shown in the oscillogram has the same nearly parabolic shape as is found by other methods. The analysis made at

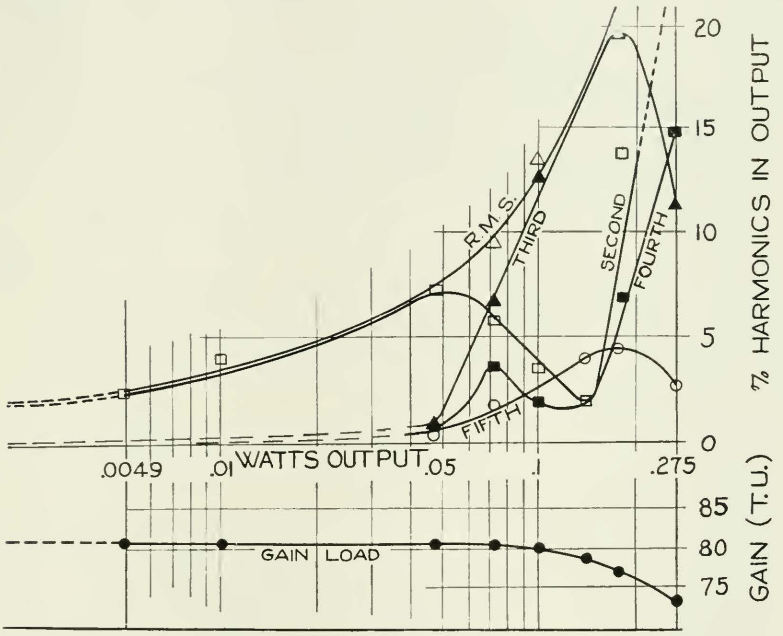


Fig. 6—Amplifier No. 1—Plate voltage 130-V, 1,000 c.p.s. input. Variation of gain and distortion with output

this point showed 7.3% second harmonic, .8% third and less than .1% fourth and fifth harmonics. The mathematical analysis of the problem expresses the $E_c - I_b$ characteristic of the tube by a power series and shows that the coefficient of the second power term in the series is the principal factor in producing second harmonic. The percentages of harmonic given therefore are such as would be obtained from a tube having a nearly parabolic characteristic.

Analyses made at lower outputs showed that the amount of second harmonic present varies with the power output in a manner described by a slightly curved line on the logarithmic scale used (Fig. 6). The third and higher harmonics are negligible at low outputs but at

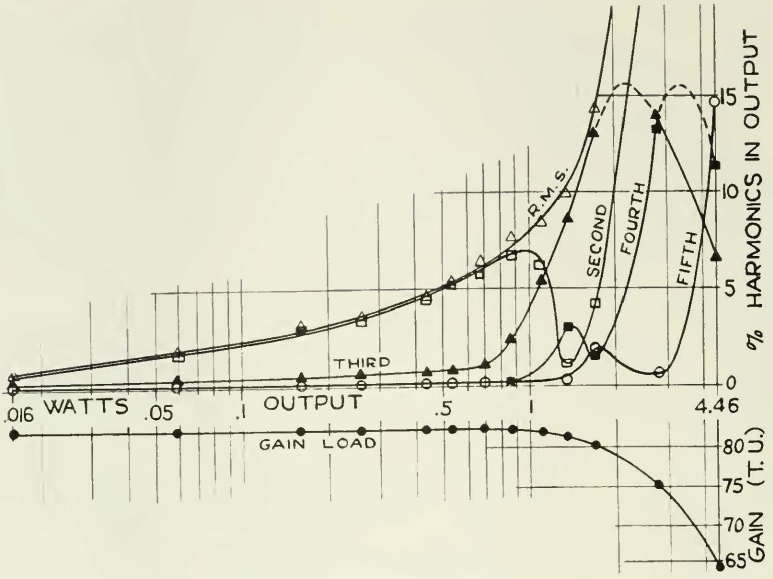


Fig. 7—Amplifier No. 1—Plate voltage 350-V, 1,000 c.p.s. input. Variation of gain and distortion with output

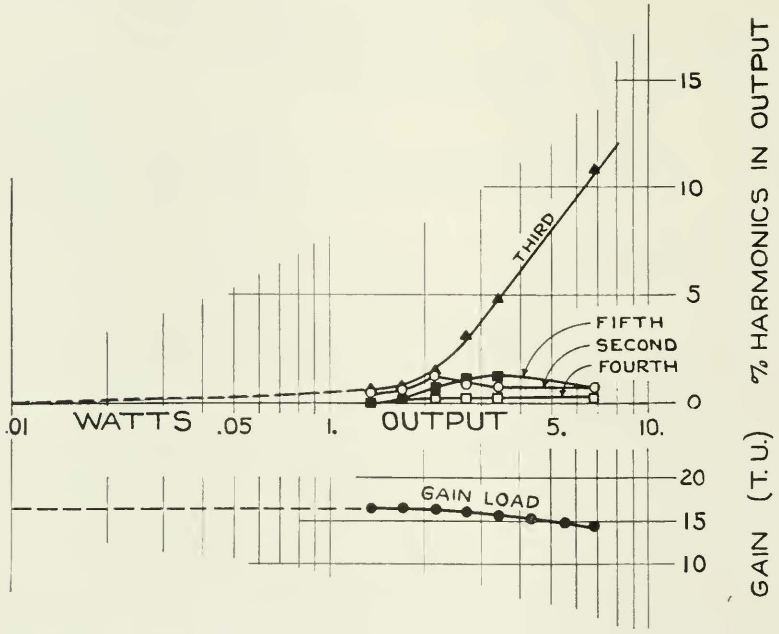


Fig. 8—Amplifier No. 2, 1,000 c.p.s. input. Variation of gain and distortion with output

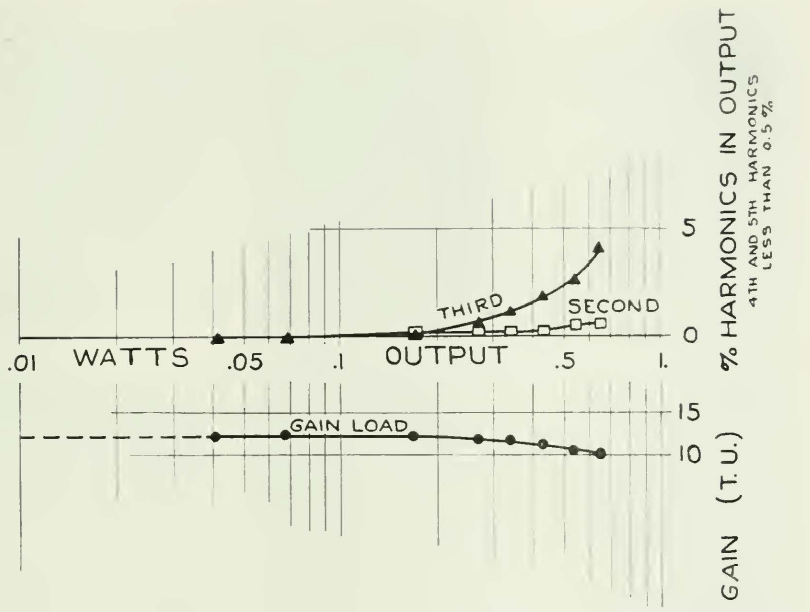


Fig. 9—Amplifier No. 3, 1,000 c.p.s. input. Variation of gain and distortion with output

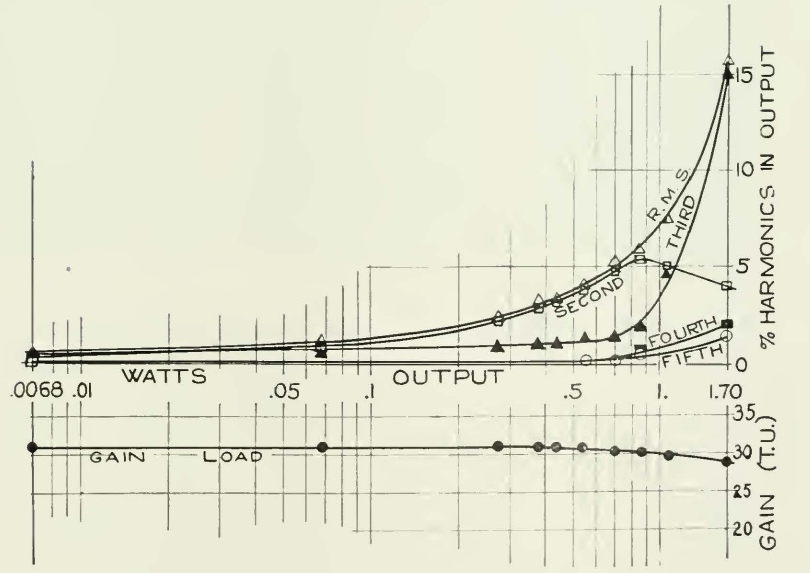


Fig. 10—Amplifier No. 4, 1,000 c.ps. input. Variation of gain and distortion with output

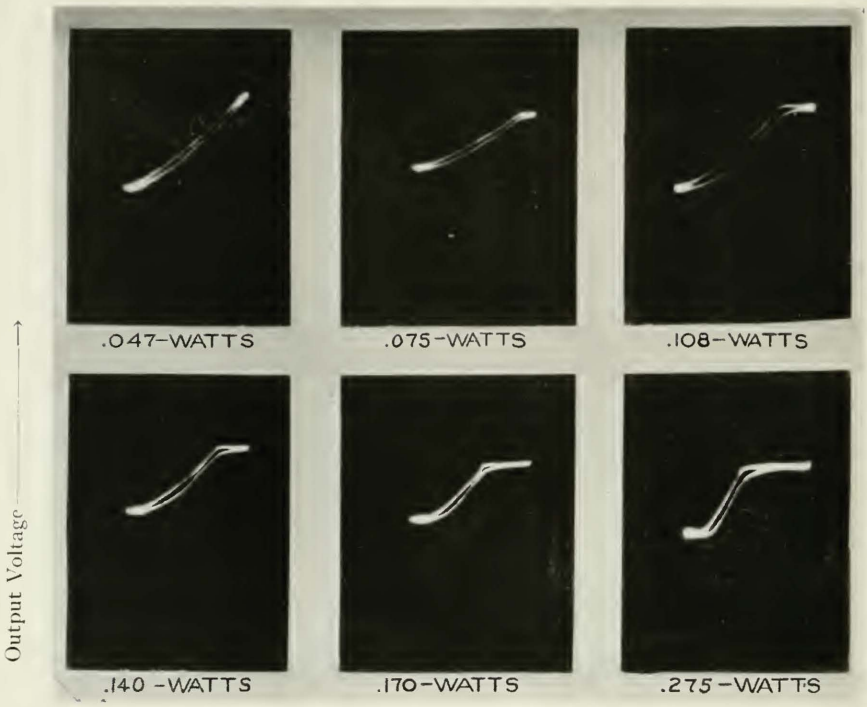


Fig. 11—Amplifier No. 1; 1,000 c.p.s.; load 4,000 ohm resistance, 130 volt plate supply

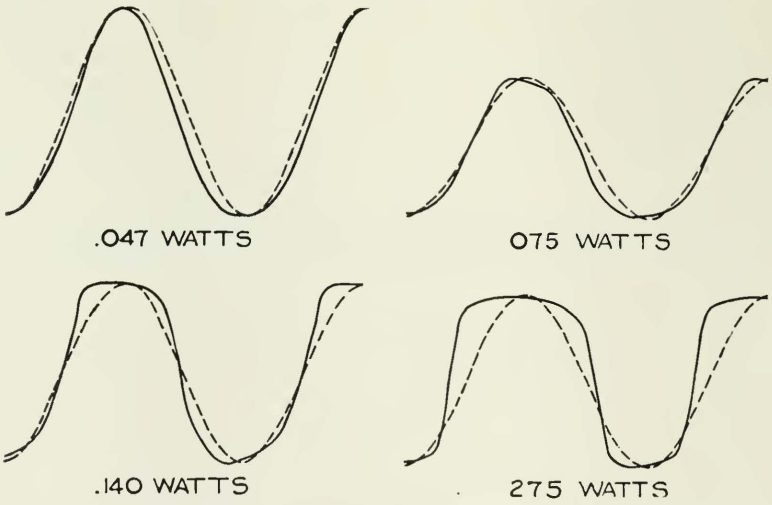


Fig. 11-A—Projection of output wave from oscillograms of Fig. 11

outputs above .05 watts the third harmonic increases rapidly and then falls somewhat while the second harmonic falls to a minimum of about 2% at .14 watt output and then increases rapidly. The fourth and fifth harmonics follow similar cycles of increase and decrease, the fourth following the second and the fifth the third.

To assist in interpreting the stages in the tube overloading at which these changes in the percentages of the different harmonics

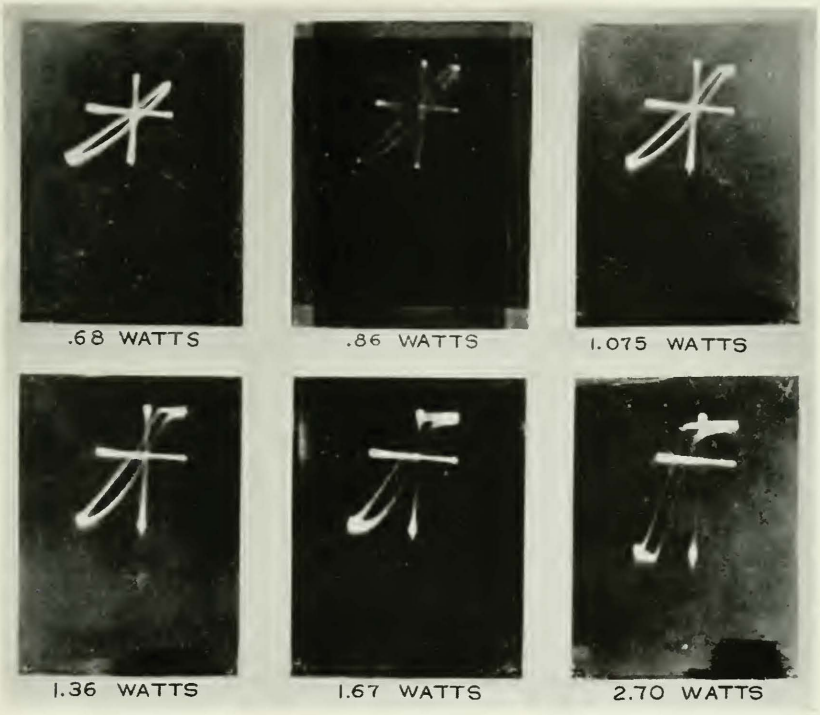


Fig. 12—Amplifier No. 1; 1,000 c.p.s.; load 4,000 ohm resistance, 350 volt plate supply

take place the waves corresponding to the vertical deflections of some of the oscillograms have been projected against a time axis by an appropriate geometrical construction, assuming the horizontal deflection a sine wave as it very nearly was.

These projections for four of the oscillograms of Fig. 11 are shown in Fig. 11-A, a pure sine wave being drawn against each figure for purposes of comparison. Up to an output of .047 watt the curved characteristic of the tube results in the asymmetrical wave shown

with a predominant second harmonic. At a point slightly above .047 watt one or more of the grids becomes positive to the filament for part of a cycle and draws current. On account of the high impedance of the circuits supplying the grids this immediately results in a flattening of the top of the wave which is well developed at .075 watt output. This flattening of the top of the wave compensates to some extent for the curvature of the lower part of the tube character-

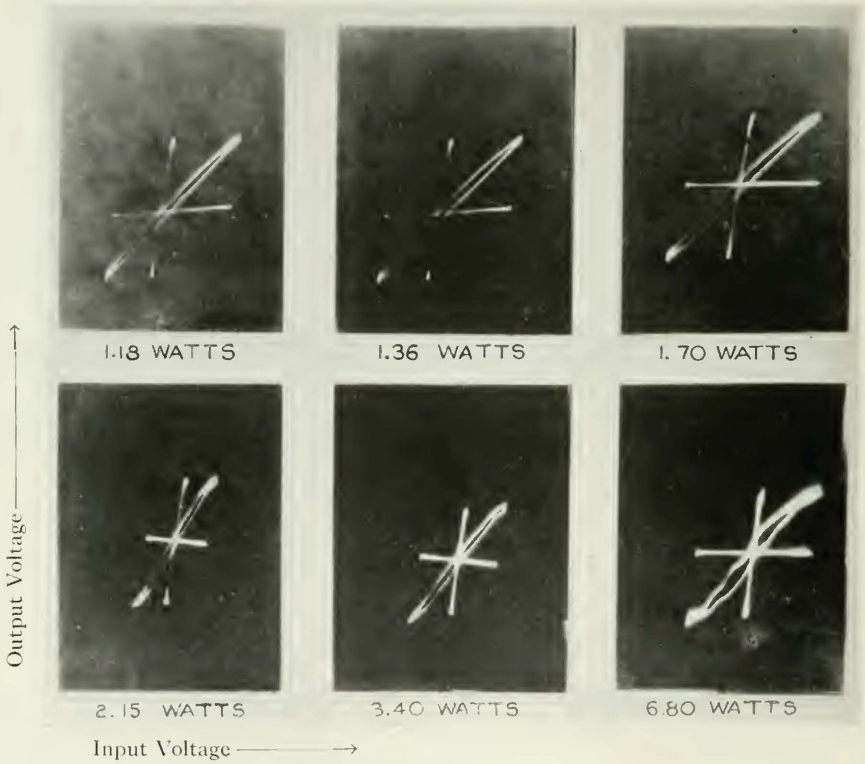


Fig. 13—Amplifier No. 2; 1,000 c.p.s., load 500 ohm resistance

istic so that the curve becomes more symmetrical with regard to the zero axis although more distorted. This corresponds to the fall of the second harmonic and increase of the third. At an output of .140 watt this compensation is more nearly complete than at any other output. At still higher outputs the top of the wave is still more flattened and the plate current is reduced to zero for a considerable part of the cycle so that the output wave becomes nearly rectangular as shown. This corresponds to large amounts of all harmonics.

In selecting from these results some point to be taken as the maximum carrying capacity of this amplifier, the question arises as to how much the second harmonic which is present at practically all loads will be noticeable. This, of course, depends on the training of the ear of the observer and on the quality of the music being transmitted. Comparing the note obtained from a cone type loud speaker when the wave corresponding to an output of .043 watts was applied with

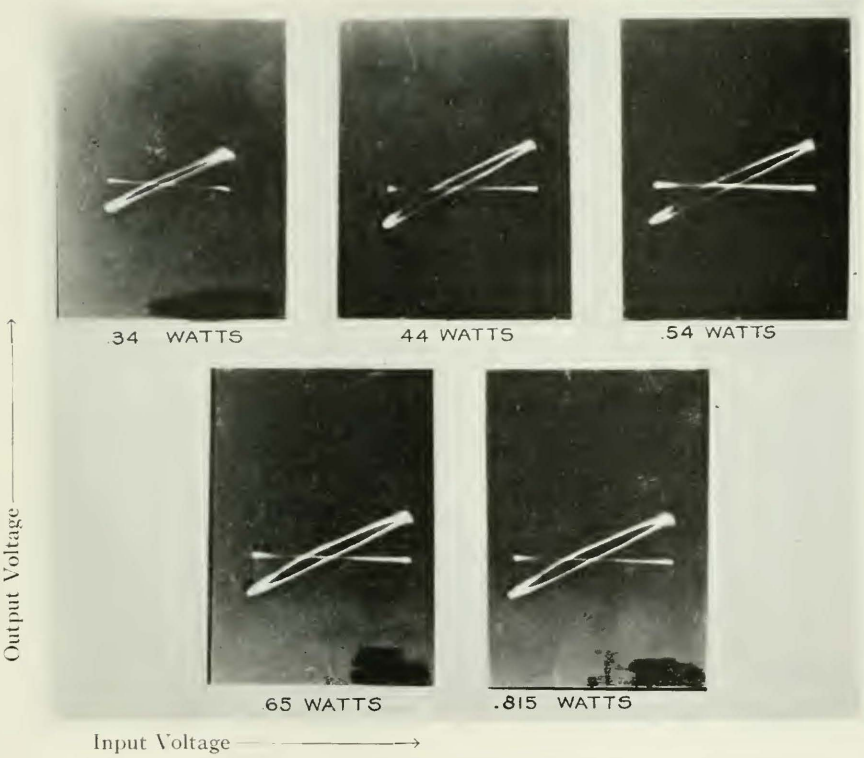


Fig. 14—Amplifier No. 3; 1,000 c.p.s.; load 600 ohm resistance

that obtained when the 1,000 cycle input was applied so as to produce a note of equal volume, it was found that it required a fairly sensitive ear to note the change in quality. On the other hand when a wave corresponding to the output wave obtained at a level of .068 watt was used for comparison the difference in quality between the pure tone and the distorted output was very easily noticed. From these data it may be assumed that when this amplifier is used in a system for the transmission and reproduction of speech or music it

is fully loaded at an output of .04–.05 watt, representing the point where the grid of a tube begins to draw current and the third harmonic increases rapidly. It will be noted from the gain-load curve that the output has to increase to .1 watt before there is a noticeable falling off in the gain of the amplifier.

For the amplifier No. 1 under the condition of 350 volt plate supply the curves and oscillograms shown in Figs. 7 and 12 are of the same

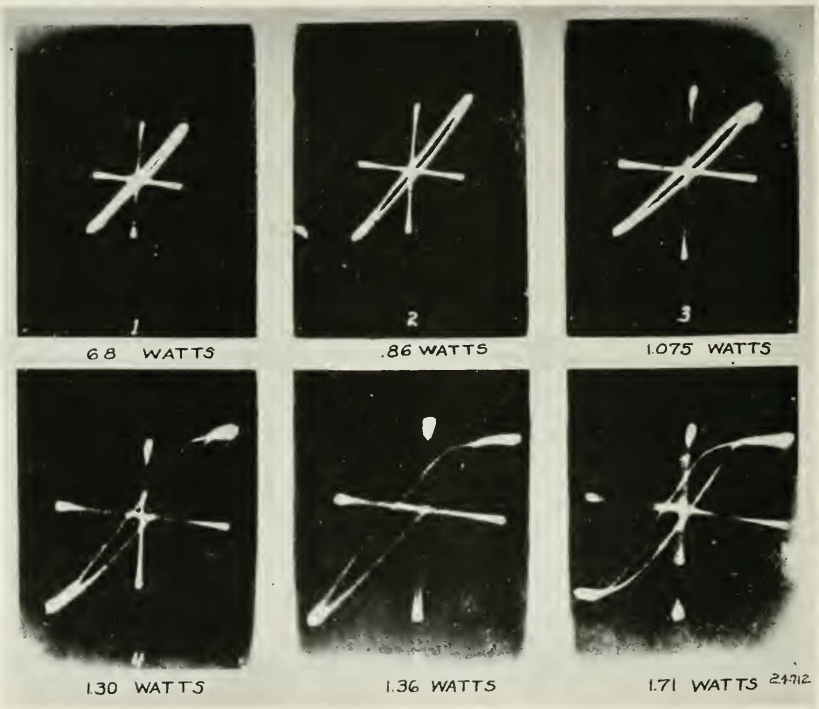


Fig. 15—Amplifier No. 4; 1,000 c.p.s.; load 4,000 ohms resistance

general nature. On account of the higher plate and grid biasing potentials an output of 0.7 watts is reached before the increase of third harmonic due to grid modulation occurs. This point is very definitely marked in the oscillograms for, of the pictures taken at outputs of 0.68 and 0.86 watts, the first is entirely free from this distortion, while the second where the current output is only 12.5% higher shows it clearly in the flattening at the top of the curve. On

the other hand the gain-load curve does not show any appreciable decrease of gain until the amplifier is considerably overloaded.

For high power amplifiers and those amplifiers in which it is desired to reduce distortion to a minimum the push-pull arrangement of tubes has been used because with this arrangement the even harmonics generated in the tubes are suppressed in the output circuit. That the suppression is quite effective is shown by the curves and

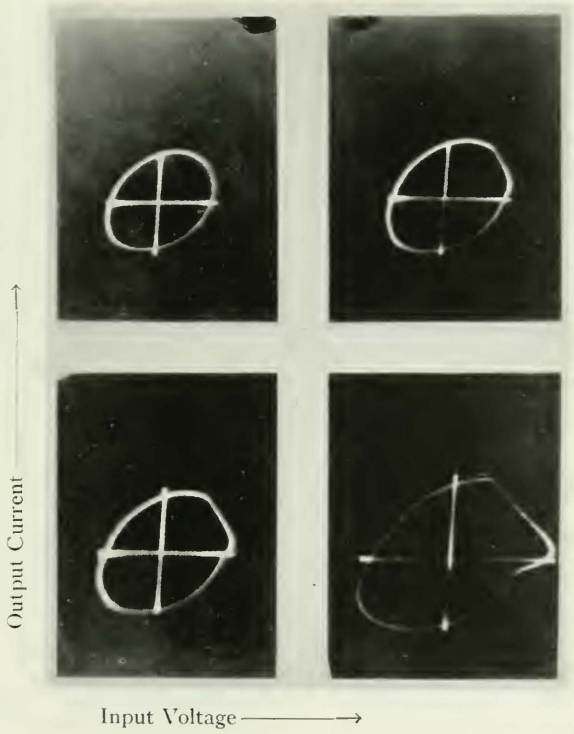


Fig. 16—Amplifier No. 4; 200 c.p.s. input; load 7,000 ohm negative reactance

oscillograms of Figs. 8, 9, 13 and 14 which were taken on the No. 2 and No. 3 amplifiers. These show that the even harmonics are very small at all loads and that the third harmonic increases suddenly at a point which in view of the plate and grid biasing potentials employed may be taken as the point at which grid modulation commences. On the oscillograms this point is not so clearly marked as in the previous cases but on those for No. 2 there is a slight flattening of the ends of the curve which is noticeable at 1.7 watts but not at 1.36 watts. On No. 3 amplifier where the impedance of the circuit

supplying the grids is lower than in No. 2 the effect of grid modulation is still less marked on the oscillograph and the rise of third harmonic in the analysis is less rapid. This amplifier was designed for an output of .365 watts. The results show that this is obtained at the expense of the introduction of about 2% third harmonic.

The No. 4 amplifier is equivalent to the last stage of the No. 1 amplifier both using about 350 volts anode potential and 27 volts

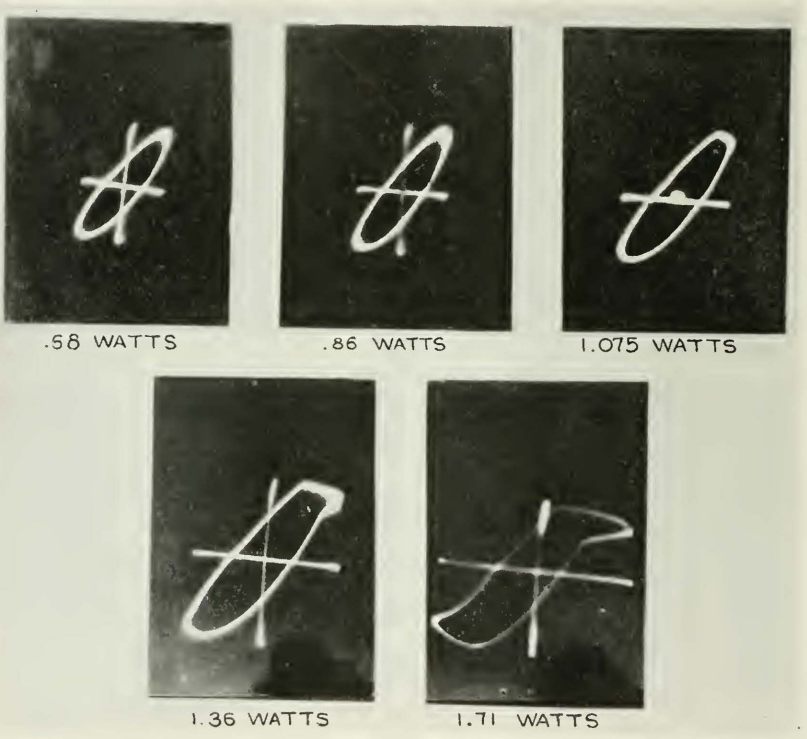


Fig. 17—Amplifier No. 4; 200 c.p.s. input; load 4,000 ohm resistance

grid bias which in the case of the No. 4 amplifier is obtained from a resistance drop in the anode circuit. The harmonic analysis curves shown in Fig. 10 indicate somewhat less second harmonic at low outputs which is probably due to the smaller number of stages. Third harmonic is approximately the same in the two cases. The oscillograph figure shows a somewhat larger output before grid modulation takes place but the difference is not great.

To check whether similar results would be obtained at other frequencies and with reactive loads oscillograms were taken with output impedances having large positive and negative phase angles at frequencies of 200 c.p.s. and 1,000 c.p.s. While the width of the ellipse obtained varied greatly as was to be expected, it was found that the points where marked irregularities in the figures occurred were at the same grid excitations as in the case of the figures taken at 1,000 c.p.s.

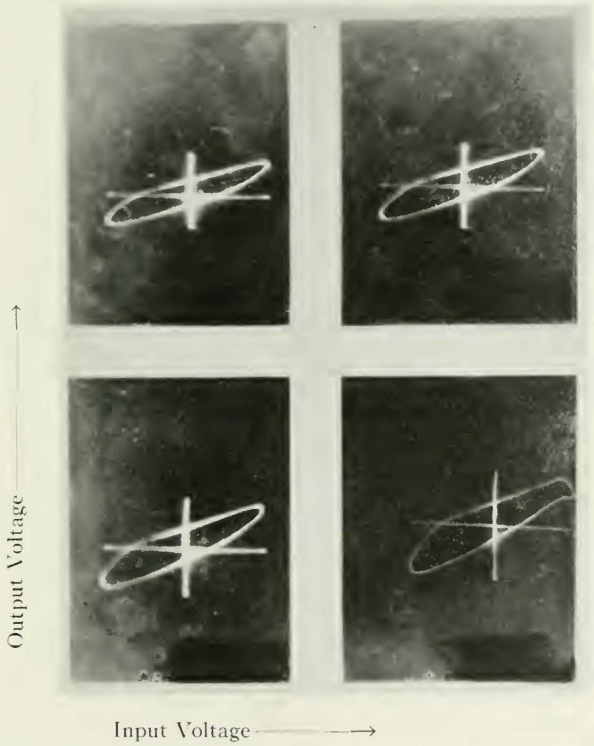


Fig. 18—Amplifier No. 4; 1000 c.p.s.; load 7,000 ohm positive reactance

with resistance load. Figures were also taken using the output current to deflect the electron stream magnetically with the same results. Typical oscillograms for these varying conditions are reproduced in Figs. 16, 17 and 18.

In conclusion, the load carrying capacity of an amplifier may be determined by either method with approximately the same results. An harmonic analysis reveals in detail the amount of each harmonic that is introduced at any load and gives useful data for fundamental

studies. It requires considerable apparatus and is slow in operation. The fall of gain method while it gives approximately the same results as the other methods is not so precise since the gain falls very slowly at the overload point and does not begin to fall rapidly till the amplifier is heavily overloaded. It is therefore, difficult to pick the exact point where overloading occurs. Moreover the method affords no indication of the kind of overloading that is occurring.

Determination by the use of the cathode ray oscillograph is more rapid and in most cases more precise although in the case of push-pull amplifiers with low grid-circuit impedances the overload point is not so clearly marked as in the other cases. The shape of the curves affords valuable information as to the place in the circuit where the overloading occurs and, by comparison with previously made analyses, a good indication of the amount of harmonic introduced. It therefore forms a very valuable tool for the design engineer.

By either method the result obtained shows the load carrying capacity of an amplifier for a single frequency. The complete answer as to how much volume in speech or music a particular amplifying system will handle depends upon an analysis of the power in the speech or music such as that given in C. F. Sacia's paper on Speech Power and Energy in the October, 1925, issue of this Journal.

BIBLIOGRPAHY

1. "Physical Measurements of Audition and Their Bearing on the Theory of Hearing," Harvey, Fletcher, *Bell System Technical Journal*, Oct., 1923, Vol. II, p. 145.

2. "Speech Power and Energy," C. F. Sacia, *Bell System Technical Journal*, Oct., 1925, Vol. IV, p. 627.

3. "A Theoretical Study of the Three-Element Vacuum Tube," John R. Carson, Proceedings I. R. E., Vol. VII, No. 2.

4. "Operation of Thermionic Vacuum Tube Circuits," F. B. Llewellyn, *Bell System Technical Journal*, July, 1926, Vol. V, p. 433.

5. "Design of Non-Distorting Power Amplifiers," E. W. Kellogg, *Journal A. I. E. E.*, May, 1925, Vol. 44, p. 490.

6. "The Performance of Amplifiers," H. A. Thomas, *Journal I. E. E.* (London), Feb., 1926, Vol. 64, p. 253.

7. "Selecting an Audio-Frequency Amplifier," D. F. Whiting, *Bell Laboratories Record*, June, 1926.

Quality Control Charts[1]

By W. A. SHEWHART

IRRESPECTIVE of the care taken in defining the production procedure, the manufacturer realizes that he cannot make all units of a given kind of product identical. This is equivalent to assuming the existence of non-assignable causes of variation in quality[2] of product. Of course, random fluctuations in such factors as humidity, temperature, wear and tear of machinery and the psychological and physiological conditions of those individuals engaged in carrying out the manufacturing procedure may give rise to some of these apparently uncontrollable variations. Knowing this, the manufacturer contents himself with trying to produce a product which is uniform and controlled—one which does not vary from one period to another by more than an amount which may be accounted for by a system of chance or non-assignable causes producing variations independent of time.

To make clear the significance of the terms "assignable causes" and "non-assignable causes," we may make use of the following illustration. Suppose a person were to fire one hundred rounds at a target. We know what probably would happen—the individual would not hit the bull's-eye every time. Possibly some of the shots would fall within the first ring, others within the second ring, and, in general, the shots would be distributed somewhat uniformly about the center of the target. We have a more or less definite picture of some of the possible reasons why the individual would not hit the bull's-eye every time, but we probably cannot assign the reasons or causes for his missing the bull's-eye in any particular instance—the causes of missing are non-assignable. Suppose, however, that the individual tended to shoot to the right of the bull's-eye. Naturally we would conclude that there was some discoverable cause for this general tendency, i.e., we would feel that the observed effect could be assigned to some particular cause.

The reason for trying to find assignable causes is obvious—it is only through the control of such factors that we are able to improve the product without changing the whole manufacturing process. But it would be a waste of time to try to ferret out or assign some cause for a

[1] A brief description of a newly developed form of control chart for detecting lack of control of manufactured product.

[2] Quality is some function of those characteristics X, Y, Z . . ., required to define a thing. For our present purpose we shall consider that quality is a function of a single characteristic X.

fluctuation in product which is no greater than that which could have resulted from the non-assignable causes as it would be to try to find the exact manner in which each of the causes contributed to missing the bull's-eye in the analogous case of target practice just considered.

Here then is the practical commercial problem—When do the observed differences between the product for one period and that for another indicate lack of control due to assignable causes, and when, on the other hand, do the differences in quality of manufactured product observed from one period to another indicate only fortuitous, chance or random effects which we cannot reasonably hope to control without radically changing the whole manufacturing process? We shall outline a typical example of the way this question arises, outline the basis for its solution and present the results in the form of a control chart.

Typical Example

Fig. 1 shows the frequency polygon for 15,050 instruments inspected for quality X. These instruments were selected at random throughout the year from a product manufactured in quantities of approximately

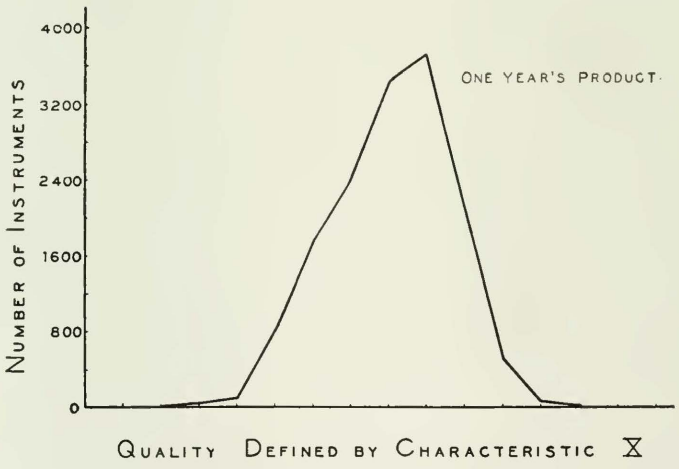


Fig. 1—Polygon showing distribution in quality for 15,050 units of product. Do these data present any evidence of lack of control?

2,000,000 per year. Is there any indication from these data that the product had not been uniform or controlled throughout the twelve month period in which the instruments had been selected?

Oftentimes we must decide from a study of a single frequency polygon of data such as given in Fig. 1, whether or not the product has been

controlled during the period for which the data have been collected. In this instance, however, it was possible to group the 15,050 observations into twelve groups representing monthly samples of approximately 1250 instruments each. The data are presented in this form

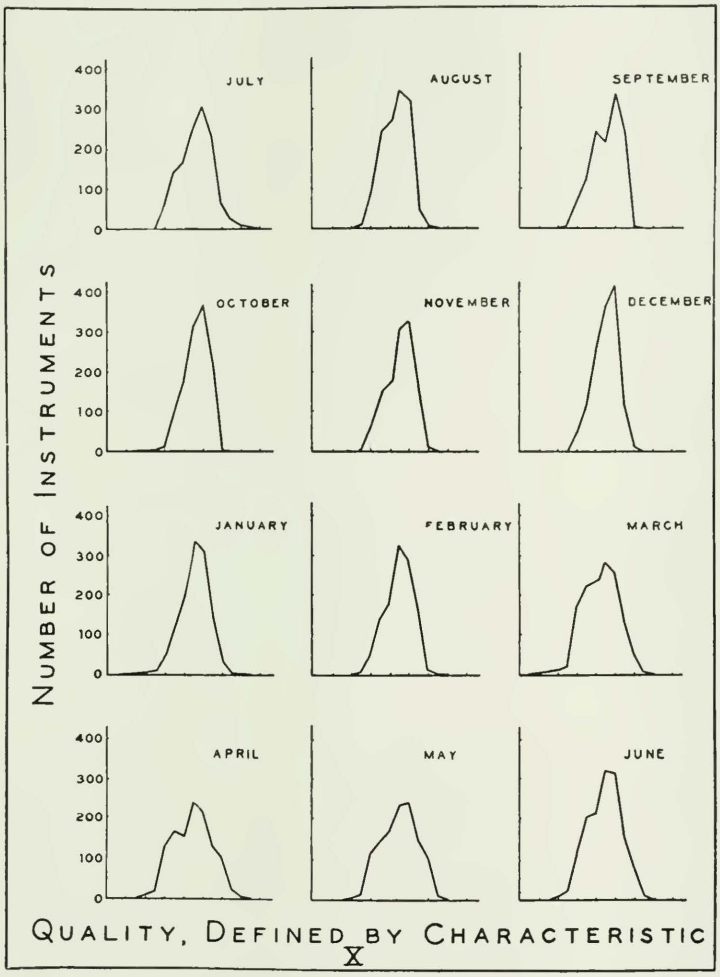


Fig. 2—Monthly polygons showing distribution in quality for samples of approximately 1250 units of product. Do these data present any evidence of lack of control?

in Fig. 2. Obviously no two polygons are the same in respect to average, dispersion and shape, but of course we would not expect them to be the same even though the product were uniform, any more than we would expect two targets to show the same distribution of shots even

if the same individual had fired at both targets. In other words, non-assignable, fortuitous or chance causes introduce certain differences in the average, dispersion and shape of the observed polygons from one month to another, and we must set up some method of differentiating the effects of assignable from those of non-assignable causes.

Outline of Basis for Detecting Lack of Control

Uniform product was defined above as one for which the differences between the units or groups of units were controlled by a complex system of non-assignable chance causes producing results independent of time. Now, following a line of reasoning whose origin is attributed to Laplace, it may be shown that such a system of causes, in general, may be expected to give a unimodal distribution of product such that the probability $dy_{\lambda'}$ of the production of a unit having the quality X within the range X to $X+dX$ is independent of time, being a continuous function, f', of the quality X and certain parameters. We may represent the probability symbolically by the following equation

$$dy_{\lambda'}=f'(X,\ \lambda_1',\ \lambda_2'\ \ldots\ \lambda_{m'}')dX, \tag{1}$$

where the λ's represent the m' parameters. Experimental evidence abounds in many fields of science to justify the adoption of Eq. 1 to represent the probability distribution of the effects of systems of chance causes. It is quite reasonable, therefore, to adopt this equation as a definition of uniform product and to use it as a basis for detecting lack of control.

Obviously, if we knew f' and the values of the m' parameters in Eq. 1, it would be comparatively easy to determine the limits within which the quality X or any estimate of a parameter derived from a sample of the product might be expected to vary because of chance causes. In practice, however, we know only the n observed values of quality obtained from inspecting a sample of as many units, and we do not know either the true functional relationship f' or any one of the m' parameters even though the product be uniform. We wish to find f' and each of the m' parameters, but, knowing that we cannot do this, we try to find some approximation f for the true function f' and some estimates $\theta_1,\ \theta_2\ \ldots\ \theta_m$ for the parameters $\lambda_1,\ \lambda_2\ \ldots\ \lambda_m$ occurring in f. To do this we tentatively assume that the sample of n units has been drawn from a uniform product distributed in accord with the function f, and then use statistical theory to see if our assumption is justified.

Theoretically there are four fundamental steps in the procedure outlined above. They are:

1. *The Problem of Specification:* To find or specify a satisfactory form f of the distribution of the uniform product from which the sample of n pieces is assumed to have been drawn or to find the equation

$$dy_\lambda = f(X, \lambda_1, \lambda_2, \ldots \lambda_m)dX \tag{2}$$

where dy_λ is the assumed probability of a unit having a quality X within the interval X to $X+dX$.

For example we often assume the distribution to be normal so that Eq. 2 becomes

$$dy_\lambda = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-m_1)^2}{2\sigma^2}} dX, \tag{2'}$$

Here $m=2$, and λ_1 and λ_2 are respectively the arithmetic mean m_1 and the root mean square (or standard) deviation σ of X as defined by the normal curve Eq. 2'.

2. *The Problem of Estimation:* To find from the data given by the sample a suitable estimate for each of the m parameters in Eq. 2. These estimates of the parameters in terms of the data of the sample are often termed statistics. If we let Θ_i represent the chosen statistic for the parameter λ_i in Eq. 2, we may rewrite this equation as follows

$$dy_\Theta = f(X, \Theta_1, \Theta_2, \ldots \Theta_m)dX \tag{3}$$

as our theoretical approximation for the assumed true (Eq. 2) probability distribution.

An estimate of a given parameter may often be obtained in a number of ways by one or more methods.

In the above illustrative case of the normal law, we must estimate the two parameters m_1 and σ (Eq. 2') from the n observed values of X in the sample. Now, it is well known[3] that σ may be expressed in an indefinitely large number of ways in terms of the arithmetic means of the absolute values of the integral powers of the deviations of X defined by Eq. 2'. Estimates of σ might be obtained in terms of the corresponding means calculated from the

[3] Whittaker and Robinson, Calculus of Observation, page 182.

sample. Two such estimates familiar to all are (letting Θ_2 stand in general for an estimate of σ, the second parameter of equation 2′)

$$\Theta_{21} = \sqrt{\frac{\pi}{2}} \Sigma \frac{|X - \bar{X}|}{n},$$

and

$$\Theta_{22} = \sqrt{\Sigma \frac{(X - \bar{X})^2}{n}},$$

where the summation extends over all the X's in the sample of n and $\bar{X}$ is the arithmetic mean of these values of X.

Thus for every λ occurring in Eq. 2, we may have many ways of securing an estimate from the sample. Of these ways, which one shall we choose? Obviously, as in the case of Θ_{21}, compared with Θ_{22}, one estimate may require less labor than another in its calculation. This, however, is not always the deciding factor, because one estimate may have a larger error than another. This leads us to the third problem.

3. *The Problem of Distribution:* To determine how each of the proposed estimates of a parameter might be distributed in a sequence of samples so that we may obtain some measure of its error.

In general we desire that estimate of a given parameter which has the smallest error or highest precision. Thus, in the case of Θ_{21}, it requires a sample of $1.14n$ to give as high a precision as the estimate Θ_{22} has for a sample of size n because the ratio of the error of Θ_{21} to Θ_{22} is $\sqrt{1.14}$. Hence the economic savings effected by using the better of two estimates may be very appreciable.

Furthermore the errors of the statistics are used in establishing the limits within which observed values of the statistics calculated from different samples may be expected to lie as will be illustrated below in discussing the data of Fig. 2. Naturally such errors are used in preparing the control chart Fig. 4.

Suppose now that we have taken the three steps outlined above and found the calculated or theoretical distribution in the form of Eq. 3. What assurance have we that the observed sample could have come from such a distribution? This question leads us to the fourth problem.

4. *The Problem of Fit:* To calculate the probability of fit between the observed and theoretical distributions.

Thus, if the n observed values of X are grouped into $m+1$ cells having frequencies $n_o, n_1, \ldots n_{m+1}$ and if the calculated or theoretical frequencies in these same cells as determined from Eq. 3 are $n_{o0}, n_{10}, \ldots n_{m0}$ where $\Sigma n_i = \Sigma n_{i0} = n$, we may calculate by Pearson's method the probability P of random samples exhibiting as large or larger values of X'' than that observed in our sample where $\chi^2 = \Sigma \frac{(n_{i0} - n_i)^2}{n_{i0}}$. If the value of probability P thus found is small, we may conclude that it is highly improbable that the sample of n units of product came from uniform product of the form assumed. Of course, this theoretically does not settle the question as to whether the sample might have come from a uniform product other than that assumed, because, as we see, f is only an assumed form for f'. Practically, however, we seem justified in concluding that it is unlikely that the product is uniform if P is small, particularly since the choice of f is customarily made upon the basis of large samples. The application of this test is illustrated in connection with the discussion of the data in Fig. 3.

Practical Application of Theory

The application of the steps just outlined will be illustrated by an analysis of the data in Figs. 1 and 2 to show that the product had not been controlled for the period therein indicated. Carrying out steps 1 and 2 we conclude that the best theoretical equation representing the data in Fig. 1 is either[4] the Gram-Charlier series (two terms) or the Pearson curve of type IV for both of which the estimates of the parameters may be expressed in terms of the first four moments μ_1, μ_2, μ_3 and μ_4 of Fig. 3. These two distributions are shown in columns 10 and 14 respectively.[5] Pearson's test for goodness of fit (step 4) gives negligible results[6] (the probabilities of fit as measured by P on the chart are for practical purposes zero) in both instances, and this was taken as indicating that assignable causes of variation had entered the product. Further investigation of an engineering nature justified this conclusion.

We should not fail to note as suggested above, however, that a small value of fit technically indicates only that the chance is small that a random sample drawn from the theoretical universe (either the two-

[4] Equations for these curves may be found in Bowley's Elements of Statistics, pages 267 and 345 respectively.

[5] Bowley's table, page 303 in his "Elements of Statistics," was used in the calculation of the Gram-Charlier graduation.

[6] Corrections were applied to take account of the number of degrees of freedom, etc., in the calculation of goodness of fit.

OBSERVATIONS

CHARACTERISTIC INSPECTED (a)	QUALITY X
UNITS (b)	
TO NEAREST	
SOURCE OF DATA	

INSPECTION ENGINEERING ANALYSIS SHEET

SUBJECT

TYPE A-INSTRUMENT
1 Quality from
2 July 1923 -
3 June 1924 incl.

CALC. BY MBC	REPORT NO.
CHK. BY MSH	DATE 1-14-26
APP'D.	
SHEETS	SHEET

$f_1(x)$ Gram-Charlier (2 terms) $f_2(x)$ Pearson Type IV

CELL NO.	Quality (a) IN X (b)	CELL BOUNDARIES LOWER	CELL BOUNDARIES UPPER	DEV. IN CELLS FROM ORIGIN AT 0. -5.5 X	OBS. FREQ. y	yX	yX²	yX³	yX⁴	THEOR. FREQ. y_1	$y-y_1$	$(y-y_1)^2$	$\frac{(y-y_1)^2}{y_1}$	THEOR. FREQ. y_2	$y-y_2$	$(y-y_2)^2$	$\frac{(y-y_2)^2}{y_2}$	CELL NO.
0	-5.5	-5.75	-5.26	0	13	0	0	0	0									0
1	-5.0	-5.25	-4.76	1	10	10	10	10	10									1
2	-4.5	-4.75	-4.26	2	8.74	16	32	64	128	86	-12	144	1.674	103	-29	841	8.165	2
3	-4.0	-4.25	-3.76	3	43	129	387	1161	3483									3
4	-3.5	-3.75	-3.26	4	100	400	1600	6400	25600	253	-153	23409	92.526	229	-129	16641	72.668	4
5	-3.0	-3.25	-2.76	5	815	4075	20375	101875	509375	675	140	19600	29.037	631	184	33856	53.655	5
6	-2.5	-2.75	-2.26	6	1761	10566	63396	380376	2282256	1466	295	87025	59.362	1463	298	88804	60.700	6
7	-2.0	-2.25	-1.76	7	2397	16779	117453	822171	5755197	2662	-265	70225	26.381	2708	-311	96721	35.717	7
8	-1.5	-1.75	-1.26	8	3431	27448	219584	1756672	14053376	3679	-248	61504	16.718	3716	-285	81225	21.858	8
9	-1.0	-1.25	- .76	9	3703	33327	299943	2699487	24295383	3441	262	68644	19.949	3477	226	51076	14.690	9
10	- .5	- .75	- .26	10	2165	21650	216500	2165000	21650000	2024	141	19881	9.823	997	168	28224	14.133	10
11	0	- .25	.24	11	510	5610	61710	678810	7466910	683	-173	29929	43.820	626	-116	13456	21.495	11
12	.5	.25	.74	12	77	924	11088	133056	1596672									12
13	1.0	.75	1.24	13	15.94	195	2535	32955	428415	81	13	169	2.086	100	- 6	36	.360	13
14	1.5	1.25	1.74	14	2	28	392	5488	76832									14
15																		15
	Σ				15050	121157	1015005	8783525	78143637	15050			301.376	15050			303.441	
										NO. OF CELLS = 10		P = .00000		NO. OF CELLS = 10		P = .00000		

MOMENTS ABOUT ORIGIN $\bar{O}$

NOTE: THE ORIGIN $\bar{O}$ IS THE MID-VALUE BETWEEN THE BOUNDARIES OF THE CELL CHOSEN

$\mu_1 = \frac{\Sigma yX}{\Sigma y} = \frac{121157}{15050} = 8.050299$

$\mu_2 = \frac{\Sigma yX^2}{\Sigma y} = \frac{1015005}{15050} = 67.442193$

$\mu_3 = \frac{\Sigma yX^3}{\Sigma y} = \frac{8783525}{15050} = 583.622924$

$\mu_4 = \frac{\Sigma yX^4}{\Sigma y} = \frac{78143637}{15050} = 5192.268239$

m = UNITS (b) PER CELL = .5

$\bar{X} = \bar{o} + m\mu_1 =$ -5.500000+4.025150 = -1.474850

CORRECTED MOMENTS

$\sigma = m\mu_2'^{1/2} =$ .5 x 1.597356 = .798678

$k = \frac{\mu_3}{\mu_2^{3/2}} = \frac{-1.730023}{4.075727} =$ -.424470

$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{22.144232}{6.510387} =$ 3.401370

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{.798678}{\sqrt{15050}} =$ .00651034

$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}} = \frac{.798678}{\sqrt{30100}} =$ .00460350

$\sigma_k = \sqrt{\frac{6}{N}} = \sqrt{.00039867} =$ .0199667

$\sigma_{\beta_2} = \sqrt{\frac{24}{N}} = \sqrt{.00159468} =$ .0399334

UNCORRECTED MOMENTS ABOUT ARITH. MEAN

$\mu_1 = 0$

$\beta_1 = k^2 = (-.424470)^2 =$ .180175

$\mu_2 = \mu_2' - \mu_1'^2 =$ 67.442193-64.807314 = 2.634879

$3\sigma_{\bar{x}} =$ .0195310

$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 =$ 583.622924-1628.789457+1043.436510 = -1.730023

$3\sigma_\sigma =$ .0138105

$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_1'^2\mu_2' - 3\mu_1'^4 =$ 5192.268239-18793.356164+26224.484274-12599.963844=23.432505

$3\sigma_k =$ .0599001

$3\sigma_{\beta_2} =$ .119800

CORRECTED MOMENTS ABOUT ARITH. MEAN (SHEPPARD'S CORRECTIONS)

$\mu_2(\text{COR.}) = \mu_2 - \frac{h^2}{12} =$ 2.634879-.08333 = 2.551546

$\mu_4(\text{COR.}) = \mu_4 - \frac{1}{2}h^2\mu_2' + \frac{7}{240}h^4 =$ 23.432505-1.317440+.029167 = 22.144232

$h =$ 1

Fig. 3

term Gram-Charlier series or Pearson IV type in this case) would give as large or larger value of χ^2 than that observed. Therefore the basis for the conclusion at the end of the previous paragraph is that we have faith[7] that the customary method of taking theoretical steps 1 and 2 gives a close approximation to the true distribution of the product when it is uniform or controlled.

Turning to a study of the data grouped into monthly distributions (Fig. 2), we find additional evidence of lack of control. Naturally the monthly observed values of the four statistics, average $\bar{X}$, standard deviation σ, skewness $k = \sqrt{\beta_1}$, and kurtosis β_2 should lie within well-defined limits established by sampling theory (step 3) and shown in Fig. 4, if the product had been controlled. Furthermore, the observed values of percentage defective p (percentage of instruments having quality less than some value X) from month to month also should fall within well-defined limits. Using the grand average[8] of a statistic as the basis for establishing limits, the first five sections of the control chart in Fig. 4 were constructed. The dotted lines calculated upon the basis of a uniform sample of 1250 indicate the limits within which the different statistics should lie, if the product had been controlled. The chart shows that observed values of these statistics often fall outside their respective limits indicating, subject to limitations imposed by the method of calculation, lack of control of product.

We may go still further and, without carrying out the analysis of Fig. 3, make use of Pearson's test of goodness of fit to calculate the probability that the first two months' samples could have been drawn from the same universe (the same uniform product), then that the third month's sample could have come from the same universe as the combined samples for the first and second months, etc.[9] Obviously the values of χ^2 used as a basis for this calculation of the goodness of fit

[7] Such faith may be based upon the a priori conception that an observed difference in two values of $\bar{X}$ is the resultant effect of a large number of causes (following in the steps of Laplace, Charlier, Edgeworth, Gram, Thiele and others) and upon the experience that observed homogeneous distributions always have been fitted by some one of the well-known forms of probability curves (following in the steps of Pearson and others).

[8] Some objection may be raised to the use of the observed average as a basis for establishing the limits of a given statistic, because this observed average almost certainly would not be the true value even though the product had been uniform. In the present case, however, we are probably justified in using the observed average because previous experience based upon thousands of observations has given approximately the same values for these quantities. Rigorously, of course, we should find the standard deviations of monthly differences from the grand average and set up limits on this basis. Wherever necessary this method is followed and in fact has been carried out for the case in hand where it gives results similar to those indicated in Fig. 4.

[9] Pearson, K.P., Biometrika, vol. viii, 1911, p. 250 and vol. x, 1914, p. 85.
Rhodes, E.C., Biometrika, vol. xvi, 1924, p. 239.

should fall within well-defined limits such as indicated on the chart. Reference to the χ^2-part of the control chart, Fig. 4, shows that this test gives more conclusive evidence than any other for deciding that the product had not been controlled. As previously noted, further investigation revealed the assignable causes of lack of control. This is a

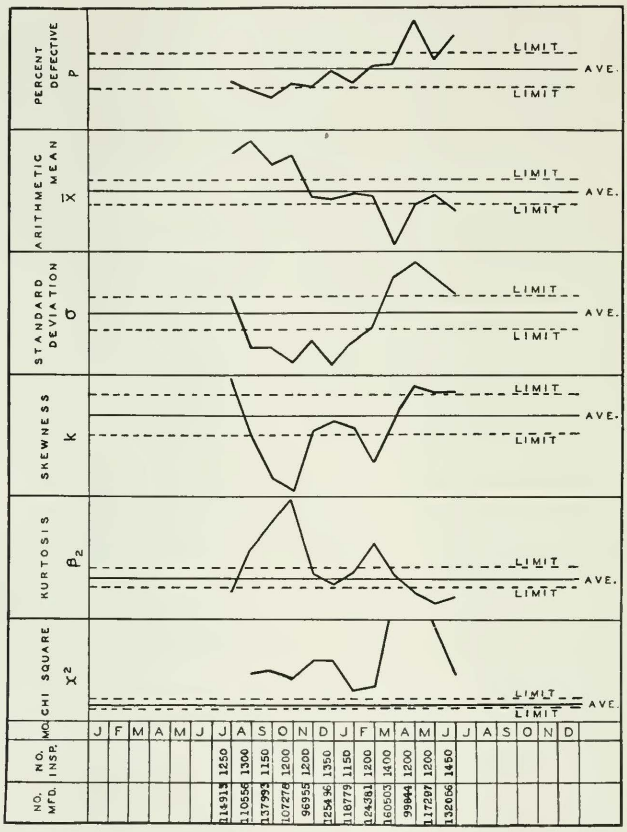


Fig. 4

common experience under such circumstances. Furthermore, it is of interest to note that the preparation of such a chart requires but a small amount of labor on the part of a computer.

Discussion and Conclusion

This paper shows how statistical methods may be used to detect lack of control of product. It describes a recently developed form of manufacturing control chart which helps in the use of inspection and pro-

duction data by applying some of the modern tools of the statistician. The chart tells the manufacturer at a glance whether or not the product has been controlled. Evidence of lack of control calls for immediate attention, but there need be no time lost in looking for causes of variation in product when these variations are not large enough to indicate lack of control.

There is an obvious advantage in using all parts of the chart wherever possible, because, as the illustration shows, one part may reveal trouble even though some other parts do not. However, when the inspection is made on the basis of attributes, the data will be available for the first or percentage defective part of the chart only.

Applications of Poisson's Probability Summation

By FRANCES THORNDIKE

SYNOPSIS: The applicability of Poisson's exponential summation to a variety of actual data is illustrated by thirty-two examples of actual frequency-distributions to which the Poisson distribution is a fairly good approximation. The comparison of actual and theoretical distributions is made graphically, using as a background new probability curves showing Poisson's exponential summation with a logarithmic scale for the average. To suggest possible explanations of the observed deviations from the theoretical Poisson distribution consideration is given to the effect on the theoretical distribution of certain modifications in the underlying assumptions, corresponding to conditions under which much actual data must be obtained.

IN an earlier number of THE BELL SYSTEM TECHNICAL JOURNAL there were published two sets of curves showing Poisson's exponential summation.[1] These charts, which are shown on a reduced scale in Figs. 1 and 2, give the relation between a, the average number of occurrences of an event in a large group of trials, the number of trials being very great compared with the average a, and the probability P that the actual number of occurrences in any such group of trials will equal or exceed any given number c. The purpose of this paper is to facilitate the use of these curves by making clear the characteristics of the Poisson summation, especially the assumptions on which it is based, and the precautions which must be observed in applying it, these points being illustrated by a number of actual frequency-distributions for which the Poisson distribution furnishes a fairly good working approximation.

POISSON'S EXPONENTIAL SUMMATION

Three assumptions underlie the mathematical treatment of Poisson's exponential summation

$$P = 1 - \left[1 + \frac{a}{1!} + \frac{a^2}{2!} + \frac{a^3}{3!} + \ldots + \frac{a^{c-1}}{(c-1)!}\right] e^{-a}$$

and its application to practical problems. The first is that the quantity measured is the number of occurrences of a particular event which always definitely happens or fails to happen, so that the actual number of occurrences c is either zero or a positive integer. The second assumption is that we may imagine the group of trials con-

[1] Figs. 1 and 2 of "Probability Curves Showing Poisson's Exponential Summation," by G. A. Campbell, *Bell System Technical Journal*, Vol. 2, No. 1, pp. 95–113, January, 1923.

stituting the sample in question to be repeated an infinite number of times, independently and uniformly, with an average number of occurrences per sample equal to a, so that we may speak of a as the average number of occurrences for the sample in question. The

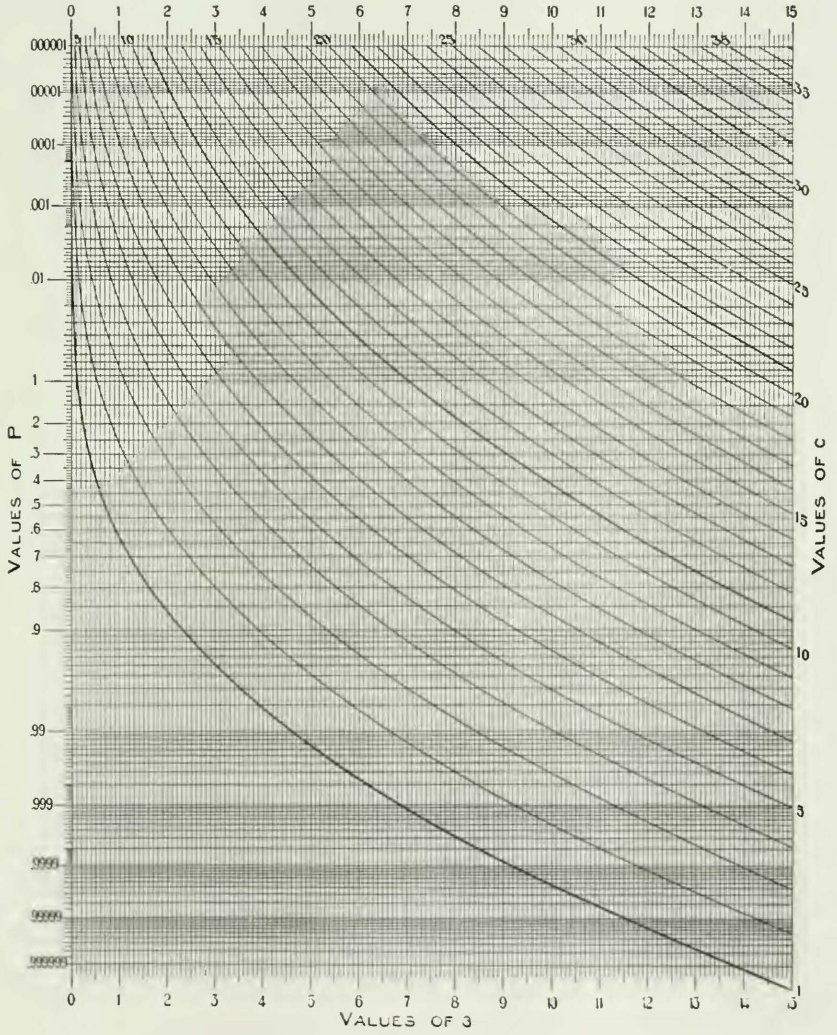


Fig. 1—Probability curves showing Poisson's exponential summation

$$P = 1 - \left[1 + \frac{a}{1!} + \frac{a^2}{2!} + \ldots + \frac{a^{c-1}}{(c-1)!}\right] e^{-a}$$

for the probability P that an event occur at least c times in a large group of trials for which the average number of occurrences is a. A scale proportional to the normal probability integral is used for P, a linear scale for a

third assumption is that, while the sample has a finite average number of occurrences, it consists of an infinite number of independent, uniform trials, so that the possible number of occurrences in a sample is infinite, and the probability that the event occur in a single trial

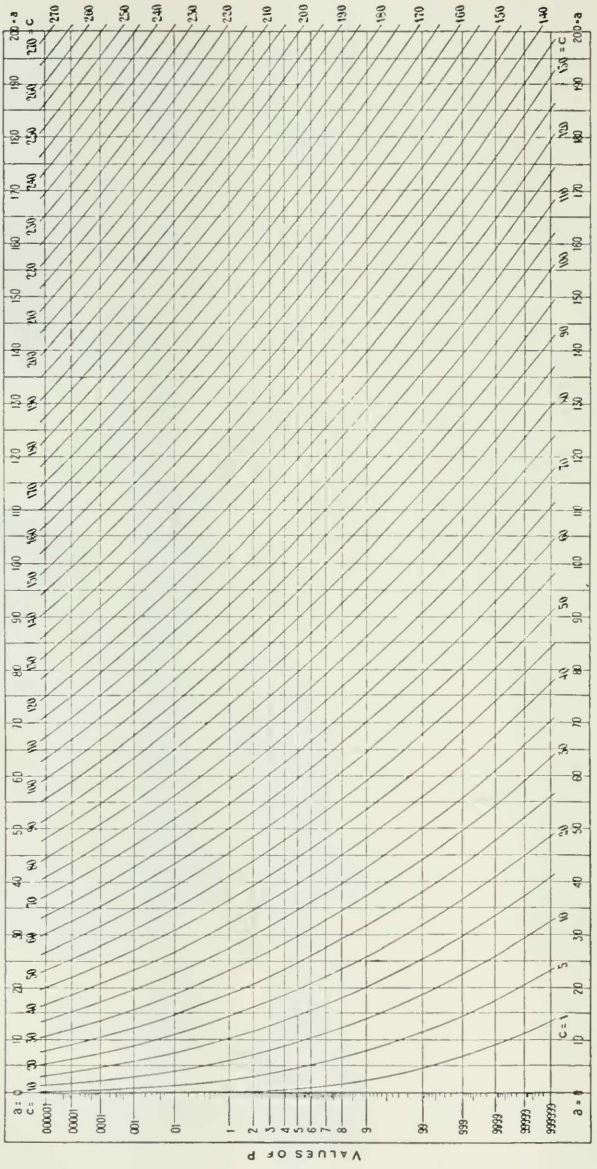


Fig. 2—Probability curves showing Poisson's exponential summation

$$P = 1 - \left[1 + \frac{a}{1!} + \frac{a^2}{2!} + \ldots + \frac{a^{c-1}}{(c-1)!}\right] e^{-a}$$

for the probability P that an event occur at least c times in a large group of trials for which the average number of occurrences is a. A scale proportional to the normal probability integral is used for P, a linear scale for a

is infinitely small. The term "uniform" applies, of course, not to the results of the trials (or samples) but to the essential conditions under which they are obtained, and "independent" is used with the meaning that the result of one trial (or sample) does not affect the occurrence of the event in any other trial (or sample). The first and third assumptions, translated into exact mathematical language, define a particular kind of probability function, which can be derived by taking the limit, as n becomes infinite and pn remains finite, of the point binomial $(p+q)^n$ for the probability of any number of occurrences of a given event in a group of n independent, uniform trials, when the probability that the event occur in a single trial is p. The second assumption is required in order that we may pass from the abstract idea of a probability function to the concrete idea of a frequency-distribution.

Throughout this discussion the summation form of the frequency-distribution, giving the probability of at least c occurrences, is used rather than the individual term form, giving the probability of exactly c occurrences. One reason for the use of the summation form is its more direct applicability to many practical problems in which the chance of exceeding a certain limit, rather than the chance of obtaining any one particular value, is of practical importance. Secondly, as Fig. 3a shows, the individual term form gives in general two possible values of c for any pair of values of a and P, whereas the summation form is single-valued and introduces no such ambiguity.

Fig. 3 also calls attention to some of the outstanding characteristics of the Poisson distribution, its discontinuity and skewness, in particular. That the Poisson distribution must be a series of discrete points and not a continuous curve is a direct result of the assumption that c represents a number of occurrences. That the distribution is skew follows from the fact that the possible number of occurrences is much larger, in fact infinitely larger, than the average number of occurrences. This skewness is quite marked even in the Poisson distribution with $a=5$, which is shown in Fig. 3, and it becomes more pronounced as a is decreased toward zero. If, for example, the average number of occurrences in a million trials is one, in any particular group of a million trials it is equally likely that there will be no occurrence of the event or one occurrence, and it is almost 1.4 times as likely that there will be no occurrence as that there will be two or more occurrences, though zero and two are equally removed from the average. A third important characteristic of the Poisson exponential, which is not brought out by this figure, is its extreme simplicity. The distribution is entirely determined by the value given to a single

parameter, the average a; its standard deviation is $\sqrt{a}$, its skewness is $1/\sqrt{a}$, and its kurtosis is $3+1/a$.[2]

One consequence of this simplicity is that there is no difficulty in deciding on a definition of the *corresponding Poisson distribution* with which any other distribution should be compared. It is naturally the Poisson distribution having the same average as the given distribution.

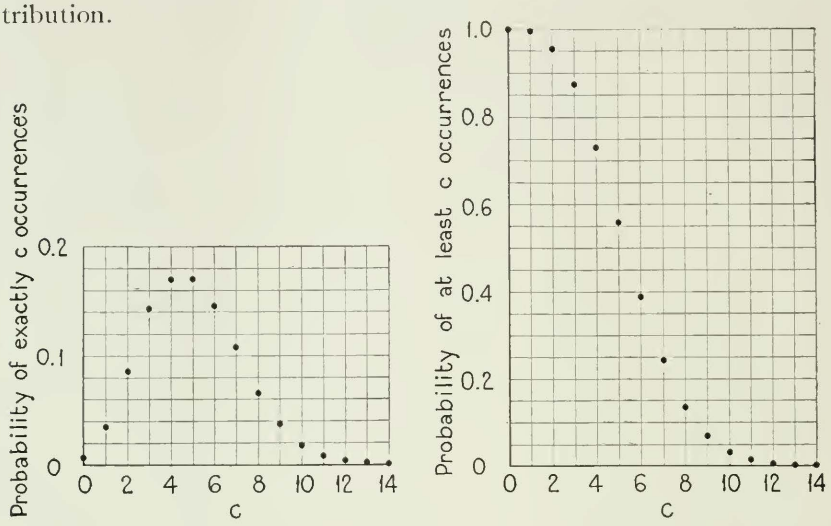


Fig. 3—Poisson distribution with the average $a=5$ shown (a) in the individual term form and (b) in the summation form

Poisson Probability Curves

Another advantage is that it is possible to represent the whole family of Poisson distributions graphically by a chart such as Fig. 1 or Fig. 2, in which the value of the average a is read on the horizontal scale, the value of the probability P on the vertical scale, and the number of occurrences c on the individual curves of the set. Any two of these three variables may then be taken as the independent variables and the values assigned to them will determine the value of the third variable, which can be read off at once. The only ambiguity occurs

[2] The standard deviation (σ), skewness (k), and kurtosis (β_2) of any distribution are defined as

$$\sqrt{\frac{\Sigma(x_i-a)^2}{N}},\quad \frac{\Sigma(x_i-a)^3}{\sigma^3},\text{ and }\frac{\Sigma(x_i-a)^4}{\sigma^4},$$

respectively, N being the number of samples in the series, and x_i the actual number of occurrences in the ith sample. For any point binomial

$$\sigma=\sqrt{npq},\quad k=\frac{q-p}{\sqrt{npq}},\quad \beta_2=3+\frac{1-6pq}{npq}.$$

when a and P are the independent variables. The point determined by their values will, in general, fall between two of the c curves and the interpretation of P must be known to determine which of the two values of c should be taken. The desired value of c is read from the lower curve if P means *a probability of P or more*, from the upper curve if P means *a probability of not more than P*.

These charts may then be used conveniently in place of unwieldy double-entry tables to obtain theoretical values needed either for comparison with experimental data or to take the place of experimental data. Examples of such uses of the Poisson exponential are discussed in detail by Karl Pearson,[3] W. A. Shewhart,[4] and E. C. Molina.[5] The use of these curves in the study of telephone trunking, letting a represent the average number of simultaneous calls from a large group of subscribers, $c-1$ the number of trunks provided for them, and P the probability that all the trunks will be in use when a subscriber attempts to make a call, is suggested by Mr. Molina's paper. Other possible applications might be found in connection with the control of errors in service, defects in a manufactured article, the stock on hand of staple articles such as ink, shoe-polish, or spark plugs, or the number of copies of reference books in a library serving a large number of people. Still others may be suggested by Table I, which is a summary of the actual data now brought together for the first time for comparison with the theory.

The comparison of any actual distribution with the corresponding Poisson distribution may easily be made graphically, using these curves as a background. In fact the charts will often be found useful as coordinate paper on which to plot any frequency-distribution, theoretical or observed, provided the values of the variate are inherently limited to the positive integers and zero.

When the curves are used in this way the corresponding Poisson distribution is represented by the points in which the vertical line for the observed value of a cuts the c curves, or for convenience simply by the vertical line itself. The other distribution may then be plotted with c and P as the independent variables, and the horizontal deviations of these points from the vertical line serve as a measure of the discrepancy between the two distributions.[6] If the comparison is to be made with an observed frequency-distribution the values used

[3] Introduction to "Tables of the Incomplete Gamma Function," London, 1922.

[4] "Some Applications of Statistical Methods to the Analysis of Physical and Engineering Data," *Bell System Technical Journal*, Vol. 3, No. 1, pp. 43–87, January, 1924.

[5] "The Theory of Probabilities Applied to Telephone Trunking Problems," *Bell System Technical Journal*, Vol. 1, No. 2, pp. 69–81, November, 1922.

[6] The distributions might be plotted in other ways, e.g., letting P or c be the dependent variable, but the method used here is the simplest.

TABLE I

N = number of samples
aN = total number of occurrences
a = average number of occurrences per sample

Series		N	aN	a
a 1	Alpha particles	2608	10097	3.87
a 2	Alpha particles	1304	10094	7.74
a 3	Deaths of aged	1096	903	0.82
a 4	Deaths of aged	1096	2364	2.16
a 5	Telephone lines in use	>1000	>4315	4.32
a 6	Bacilli	1000	1927	1.93
b 1	Yeast cells	400	720	1.80
b 2	Yeast cells	400	1872	4.68
b 3	Lost articles	423	439	1.04
b 4	Number *12*	500	421	0.84
b 5	Fires	364	9487	26.1
b 6	Incorrect reports	506	138	0.27
b 7	Cutoffs	506	1057	2.09
b 8	Double connections	506	1760	3.48
b 9	Calls for wrong number	506	2520	4.98
c 1	Deaths from kick of horse	200	122	0.61
c 2	Number *12*	250	251	1.00
c 3	Calls from group of two coin-box telephones	145	172	1.19
c 4	Calls from group of four coin-box telephones	140	384	2.74
c 5	Calls from group of two coin-box telephones	141	212	1.5[illegible]
c 6	Calls from group of six coin-box telephones	138	468	3.39
c 7	Cutoffs	267	557	2.09
c 8	Double connections	267	906	3.39
c 9	Calls for wrong number	267	1351	5.06
c10	Connections to wrong number	267	2334	8.74
c11	Party lines	300	1981	6.60
c12	"Lost and found" advertisements	209	7051	33.7
d 1	Number *12*	100	421	4.21
d 2	Number *12*	50	421	8.42
d 3	Comets	100	258	2.58
d 4	Particles in emulsion	50	46	0.92
d 5	Particles in emulsion	50	106	2.12

for the probability P are the values of the observed relative frequency F, which are calculated as indicated in Table II, and the observed distribution is represented by an irregular series of dots, as in Fig. 4.

A third set of curves, Fig. 5, supplementary to Figs. 1 and 2, has now been drawn using a logarithmic scale for a. This chart shows the individual c curves up as far as $a=30$ and it shows more clearly than does Fig. 1 the range $0.1 \leqq a \leqq 2$. It may also be used as a background in the same way as Figs. 1 and 2, with the additional advantage of making the distances of the plotted points from the vertical line proportional to the percentage deviations rather than proportional to the absolute values of the deviations, so that the fit of a distribution having a small average can be compared directly by eye with that of a distribution having a large average, since it is more often the relative than the absolute value of the deviation which is significant.

Practical Applications

In applying the Poisson summation to any concrete problem, or in comparing any observed distribution with the corresponding Poisson distribution, it is necessary to bear in mind several practical conditions which must work against any perfect agreement between the

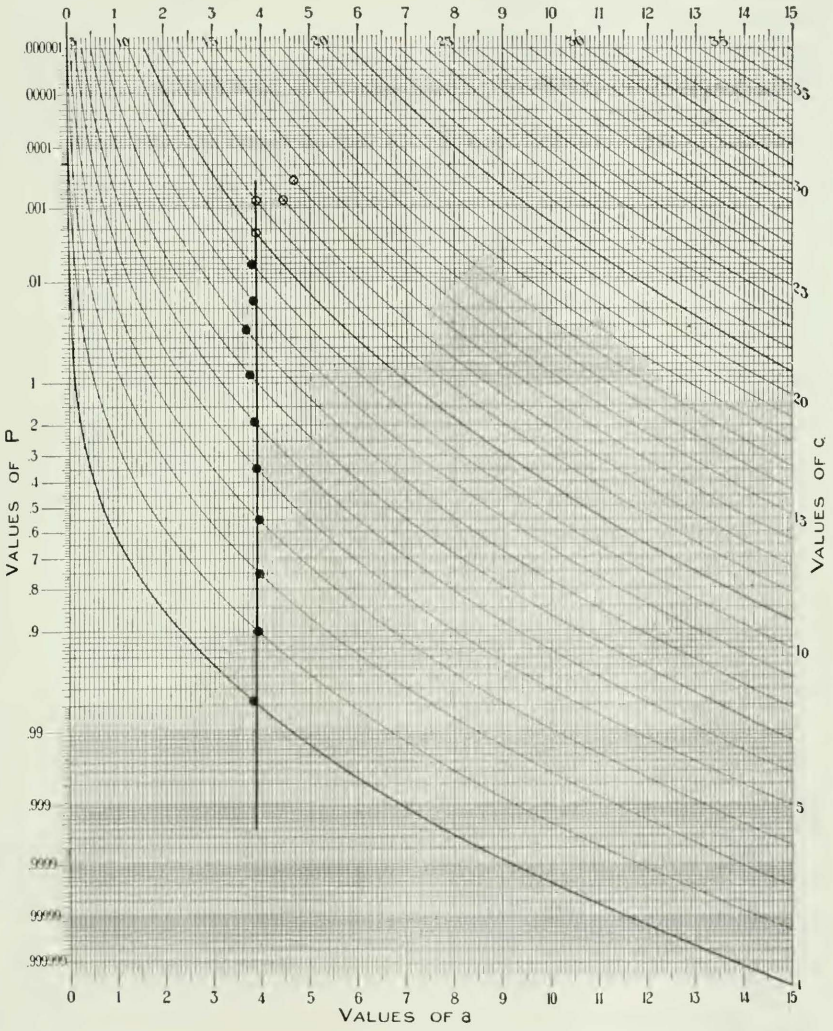


Fig. 4—Comparison of an observed distribution of the number of α particles emitted with the corresponding Poisson distribution, showing the method of using Fig. 1 or Fig. 2 as a background for plotting actual distributions. The Poisson distribution is shown by a vertical line, the observed distribution by dots

observed distribution and the corresponding Poisson distribution. In the first place, the sample considered will necessarily consist of a finite number of trials instead of an infinite number as assumed in the mathematical theory, and the trials may not be completely independent or entirely uniform. Secondly, even if the individual sample possessed the ideal characteristics assumed in the mathematical formulation, the actual series of samples must be finite and the samples may be interdependent and far from uniform. The size of the samples relating to the economic, geographic, and time divisions ordinarily used in statistical work generally varies considerably. The effect of modifying the original mathematical assumptions to correspond with some of these actual conditions is illustrated by Figs. 6–8, which show various theoretical frequency-distributions plotted on Fig. 1 or Fig. 2 for comparison with the corresponding Poisson distributions.

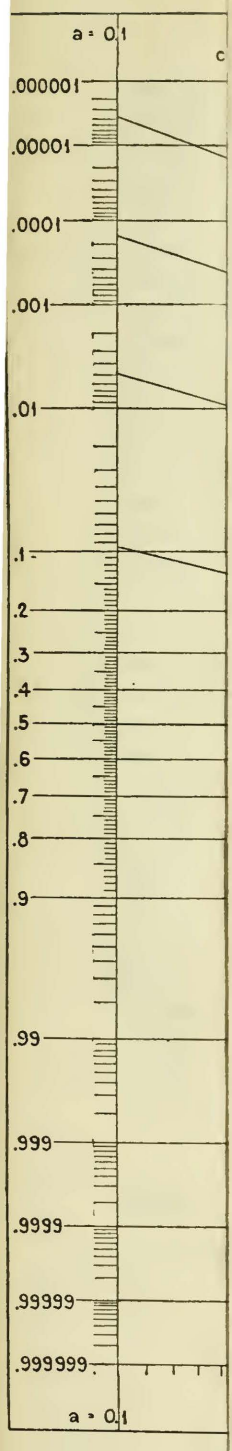
The finiteness of the number of trials n not only makes impossible the occurrence of values of c greater than the value of n, but also tends to produce a general trend away from the Poisson distribution. This is illustrated by the four typical finite binomial distributions shown in Fig. 6, which have a definite curve and slope toward the left which becomes more pronounced as n is decreased.[7] Interdependence of the trials constituting a sample will also tend to give the resulting distribution a slant, to the right if the correlation is positive, to the left if the correlation is negative.[8] Thirdly, even though the trials are independent, if they are not uniform, there will be a tendency for the distribution to slant to the left.

The requirement that N, the number of samples in the actual series, be finite introduces a somewhat different kind of deviation from the theoretical Poisson distribution. The observed relative frequency F, which is compared with the theoretical probability P, is an integral multiple of $1/N$, so that, since N is finite, the points representing the observed distribution (except those at $P=0$ and $P=1$, for which the ordinates are plus and minus infinity, and which, therefore, never appear on the graph) are all in the finite range between the two horizontal lines $P=1/N$ and $P=1-1/N$. Not only is the occurrence of points outside this range impossible, but the points near its extremes, being determined by a comparatively small number of samples, are of less significance than those near the center.

To call attention to these facts all observed distributions shown here have been represented, as in Fig. 4, with the vertical line rep-

[7] A more detailed discussion of the effect of finite sampling will be found in the paper by G. A. Campbell previously referred to.

[8] See "Explanation of Deviations from Poisson's Law in Practice," by "Student," *Biometrika*, Vol. 12, pp. 211–215, 1919.



tion terminated at
points in the range
dots and the points
centers. This sec-
ase in reliability of
d is gradual. There
ven in the center as

introduce a definite
it to the right such
when the value of a
7 shows three theo-
same average $a = 75$,
g $a = 50$ and $a = 100$,
ratio of 3:1, having
ual sub-series having
shows the effect on
l uniformly between
de up of two equal
vn for comparison.[10]
r decreases with the
to occur frequently.
material whether a
uls in the sample, or
nt's happening at a
oughout the series a
ies the tendency to
levices may be em-
ual series, some of
below.
on summation only
be some reason to

$P_i = P(c, a_i)$.

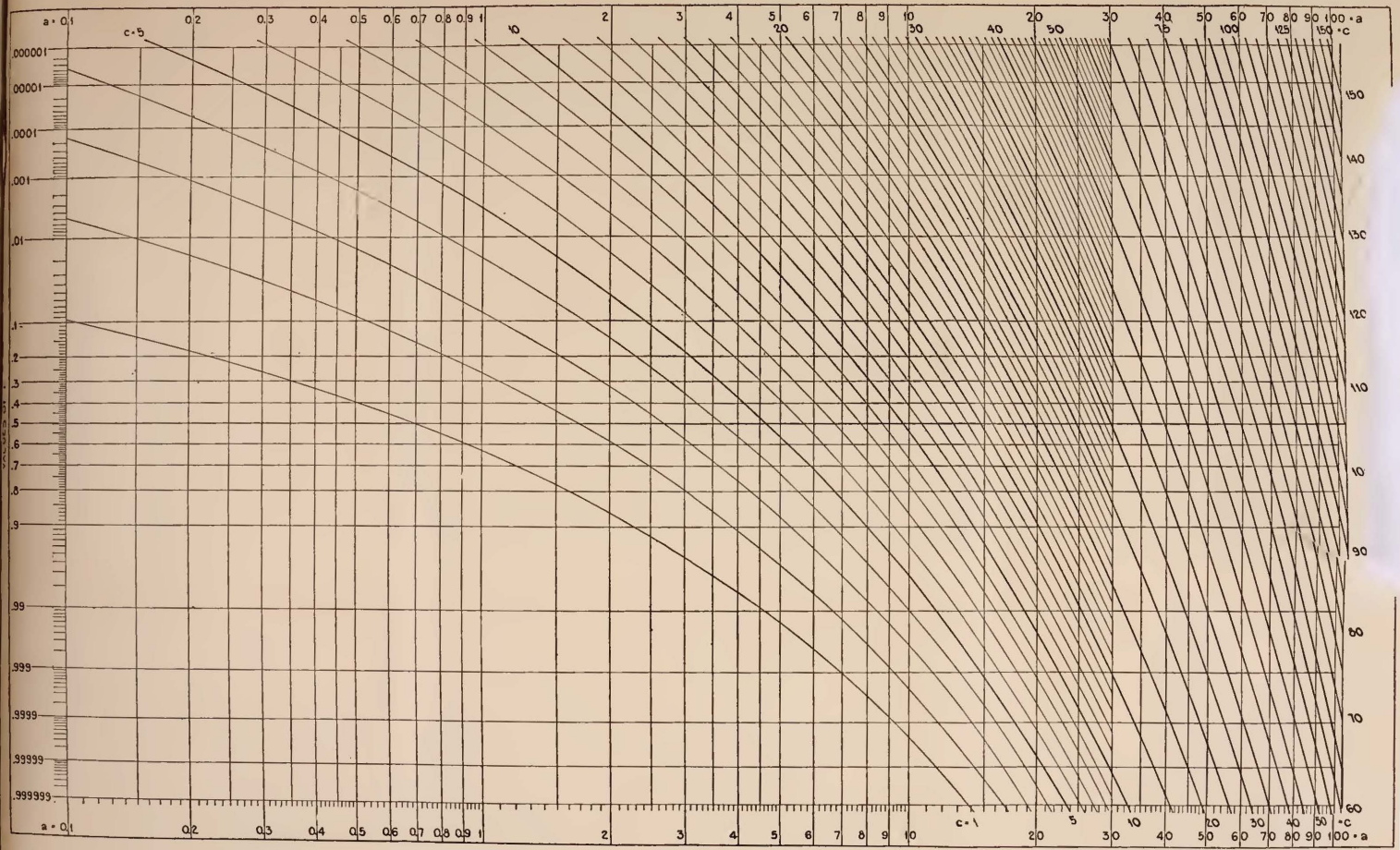


Fig. 5—Probability curves showing Poisson's exponential summation

$$P = 1 - \left[1 + \frac{a}{1!} + \frac{a^2}{2!} + \ldots + \frac{a^{c-1}}{(c-1)!}\right] e^{-a}$$

for the probability P that an event occur at least c times in a large group of trials for which the average number of occurrences is a. A scale proportional to the normal probability integral is used for P, a logarithmic scale for a

resenting the corresponding Poisson distribution terminated at $P=1\ N$ and $P=1-1\ N$, and with the observed points in the range $P=10\ N$ to $P=1-10\ N$ shown as solid black dots and the points outside this range shown as circles with white centers. This secondary division is quite arbitrary, for the increase in reliability of the points as the center of the range is approached is gradual. There will, of course, be irregularities due to sampling even in the center as long as the number of samples is finite.

Non-uniformity of the samples of the series may introduce a definite trend away from the Poisson distribution, a slant to the right such as is shown in Figs. 7 and 8. Such trends result when the value of a varies from sample to sample of the series. Fig. 7 shows three theoretical distributions of this sort, each having the same average $a=75$. Series (a) is made up of two equal sub-series having $a=50$ and $a=100$, respectively, (b) of two unequal sub-series, in the ratio of 3:1, having $a=60$ and $a=120$, respectively, and (c) of three equal sub-series having $a=15$, $a=60$, and $a=150$, respectively.[9] Fig. 8 shows the effect on the distribution of letting a vary continuously and uniformly between the limits 5 and 15, the compound series (b) made up of two equal sub-series with averages 5 and 15 being also shown for comparison.[10] Since in practical time series a usually increases or decreases with the time, this kind of distribution may be expected to occur frequently. It should be noted that in all these cases it is immaterial whether a changes because of a change in the number of trials in the sample, or because of a change in the probability of the event's happening at a single trial, or because of both; if a is constant throughout the series a Poisson distribution will be obtained, and if a varies the tendency to slope to the right will be introduced. Various devices may be employed to keep the average constant in an actual series, some of which will be illustrated by the examples given below.

In selecting the following examples of the Poisson summation only two general rules were followed: that there must be some reason to

[9] In a compound distribution

$$P=\Sigma\frac{N_i}{N}P_i$$

where N_i is the number of samples with the average a_i, and $P_i=P(c, a_i)$.

[10] If a varies uniformly and continuously from a_1 to a_2

$$P=\int_{a_1}^{a_2}\frac{P(c, a)}{a_2-a_1}da$$

$$=1-\frac{1}{a_2-a_1}\sum_{i=0}^{c}[P(i, a_2)-P(i, a_1)].$$

observed distribution
the first place, the
finite number of tri
the mathematical th
pendent or entirely
possessed the ideal
formulation, the actu
may be interdepende
relating to the econ
used in statistical w
modifying the origin
some of these actual
various theoretical f
for comparison with

The finiteness of t
the occurrence of val
to produce a general
is illustrated by the
Fig. 6, which have
becomes more prono
trials constituting a
tribution a slant, to
if the correlation is
independent, if they
distribution to slant

The requirement
series, be finite int
from the theoretic
frequency F, which
is an integral multi
representing the ol
$P=1$, for which the
therefore, never app
the two horizontal
occurrence of point
near its extremes, b
of samples, are of l

To call attention
here have been rep

[7] A more detailed dis
paper by G. A. Campbe
[8] See "Explanation of
Biometrika, Vol. 12, pp.

resenting the corresponding Poisson distribution terminated at $P=1/N$ and $P=1-1/N$, and with the observed points in the range $P=10/N$ to $P=1-10/N$ shown as solid black dots and the points outside this range shown as circles with white centers. This secondary division is quite arbitrary, for the increase in reliability of the points as the center of the range is approached is gradual. There will, of course, be irregularities due to sampling even in the center as long as the number of samples is finite.

Non-uniformity of the samples of the series may introduce a definite trend away from the Poisson distribution, a slant to the right such as is shown in Figs. 7 and 8. Such trends result when the value of a varies from sample to sample of the series. Fig. 7 shows three theoretical distributions of this sort, each having the same average $a=75$. Series (a) is made up of two equal sub-series having $a=50$ and $a=100$, respectively, (b) of two unequal sub-series, in the ratio of 3:1, having $a=60$ and $a=120$, respectively, and (c) of three equal sub-series having $a=15$, $a=60$, and $a=150$, respectively.[9] Fig. 8 shows the effect on the distribution of letting a vary continuously and uniformly between the limits 5 and 15, the compound series (b) made up of two equal sub-series with averages 5 and 15 being also shown for comparison.[10] Since in practical time series a usually increases or decreases with the time, this kind of distribution may be expected to occur frequently. It should be noted that in all these cases it is immaterial whether a changes because of a change in the number of trials in the sample, or because of a change in the probability of the event's happening at a single trial, or because of both; if a is constant throughout the series a Poisson distribution will be obtained, and if a varies the tendency to slope to the right will be introduced. Various devices may be employed to keep the average constant in an actual series, some of which will be illustrated by the examples given below.

In selecting the following examples of the Poisson summation only two general rules were followed: that there must be some reason to

[9] In a compound distribution

$$P=\sum \frac{N_i}{N} P_i$$

where N_i is the number of samples with the average a_i, and $P_i=P(c, a_i)$.

[10] If a varies uniformly and continuously from a_1 to a_2

$$P=\int_{a_1}^{a_2} \frac{P(c, a)}{a_2-a_1}\, da$$

$$=1-\frac{1}{a_2-a_1}\sum_{i=0}^{c}[P(i, a_2)-P(i, a_1)].$$

suppose the possible number of occurrences n to be at least thirty times the average a and at least 25, and that N, the number of samples in the series, must be at least 50. This last requirement excludes from our list a number of series which have previously been presented

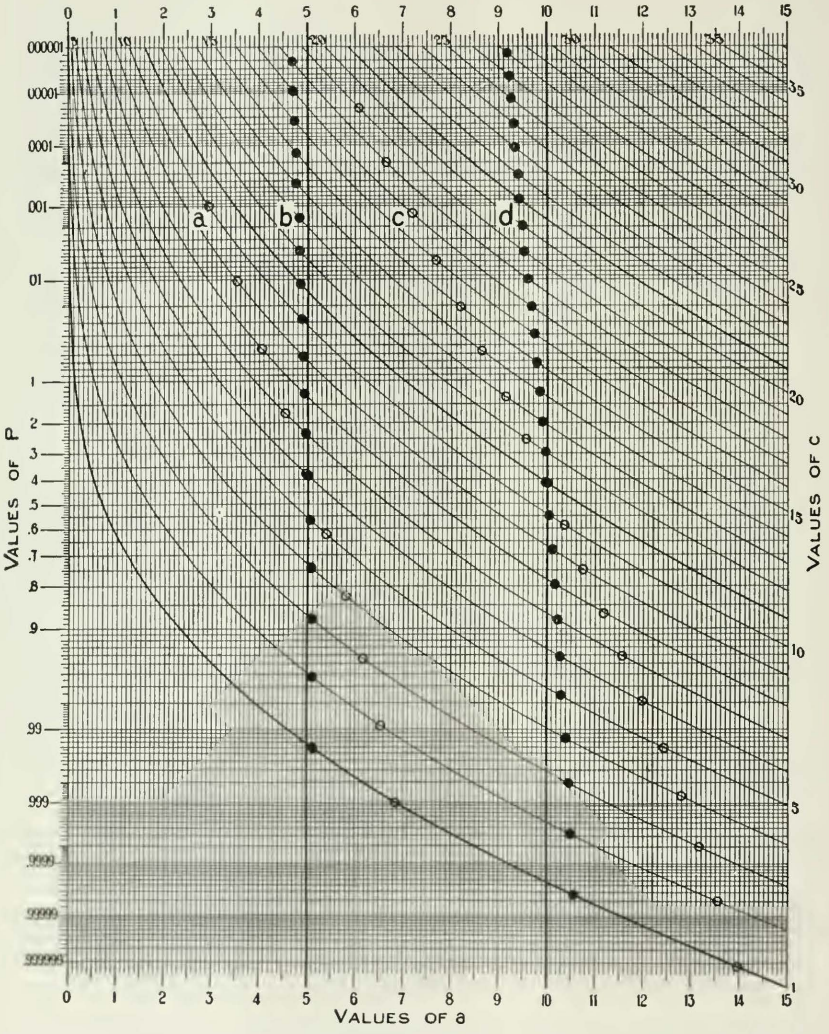


Fig. 6—Typical finite binomial distributions for the probability that an event occur at least c times in a group of n trials for which the average number of occurrences is $a = np$

(a) $a = 5,\ n = 10$
(b) $a = 5,\ n = 100$
(c) $a = 10,\ n = 20$
(d) $a = 10,\ n = 100$

as examples of the Poisson exponential, in particular those of Mortara [11] and all but one of those given by Bortkewitsch.[12]

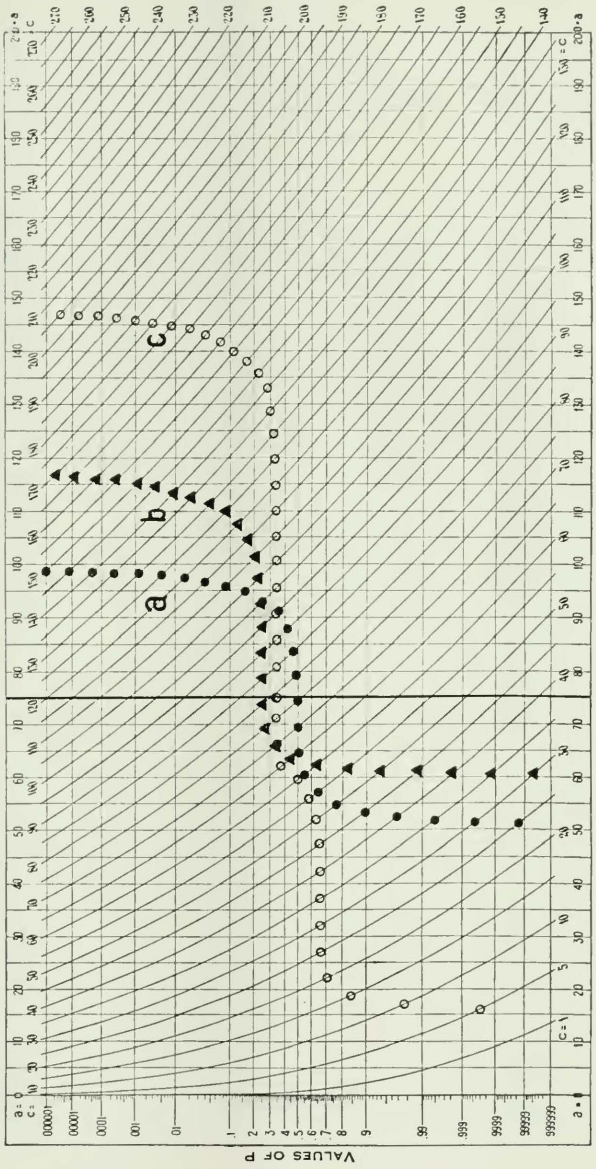


Fig. 7—Theoretical distributions for series in which
(a) $a = 50$ for one half of the samples and $a = 100$ for the other half
(b) $a = 60$ for three quarters of the samples and $a = 120$ for the other quarter
(c) $a = 15$ for one third of the samples, $a = 60$ for one third, and $a = 150$ for the other third

[11] "Sulle Variazione di Frequenza di Alcuni Fenomeni Demografici Rari," by Giorgio Mortara, *Annali di Statistica*, Series V, Vol. 4, pp. 5–61, 1912.

[12] "Das Gesetz der kleinen Zahlen," by L. von Bortkewitsch, Leipzig, 1898.

Each of the thirty-two actual distributions shown in Fig. 9 has been plotted using Fig. 5 as the background, so that the percentage deviations in all distributions may be compared directly by inspection without regard to the magnitude of the average. The examples are divided into four groups according to the number of samples in the series, and are arranged in each group roughly in order of decreasing agreement of the observed with the theoretical distributions. A

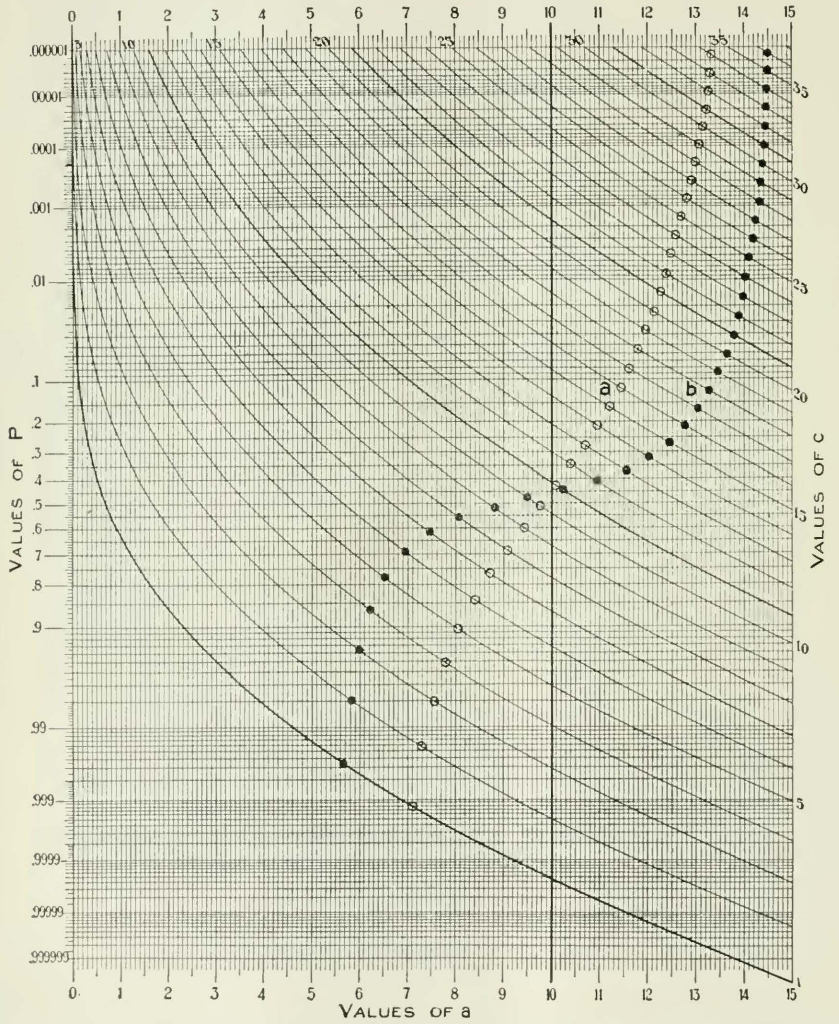
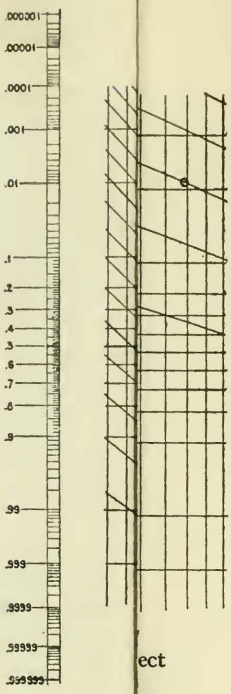
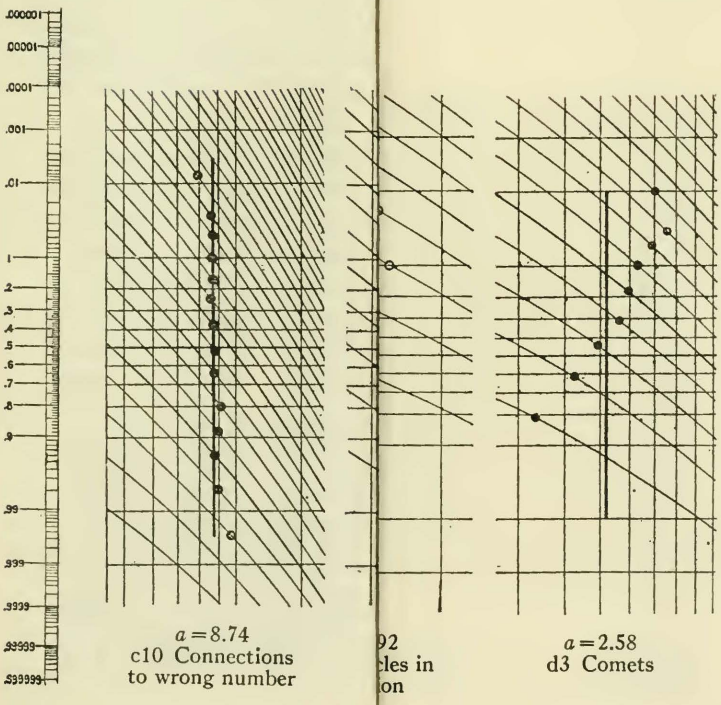


Fig. 8—Theoretical distribution for a series in which the average a varies continuously and uniformly from 5 to 15



ect



$a = 8.74$
c10 Connections
to wrong number

92
les in
on

$a = 2.58$
d3 Comets

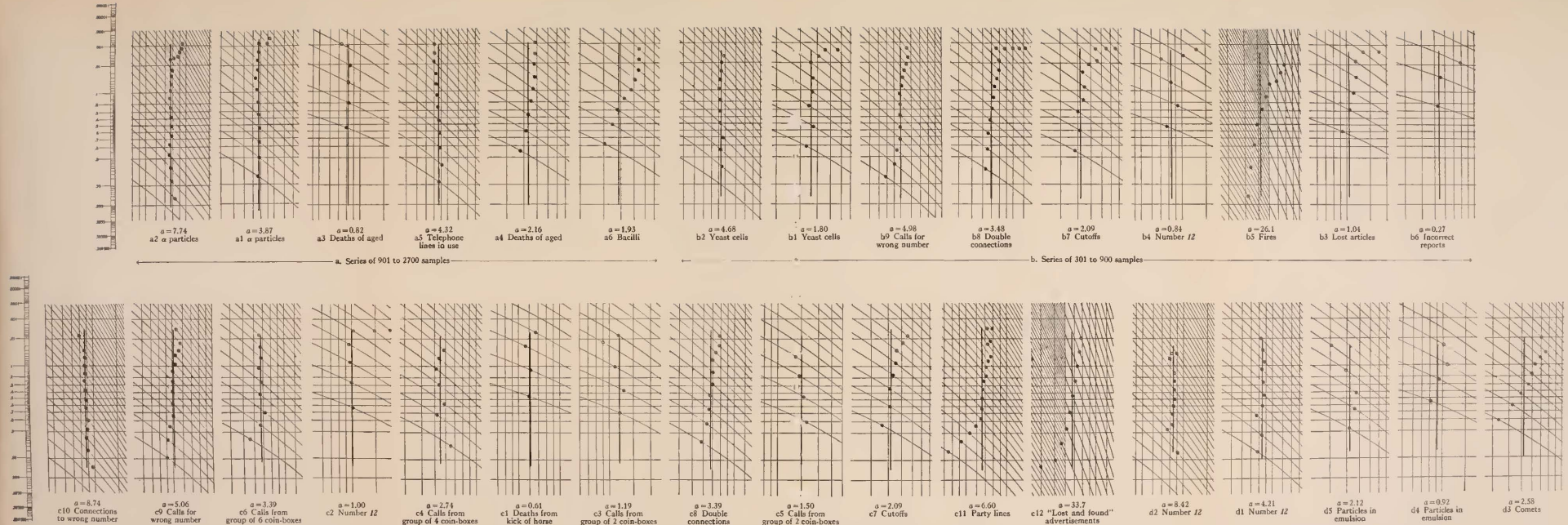


Fig. 9—Comparison of observed distributions with the corresponding Poisson distributions, using Fig. 5 as a background

summary of the data used is given in Table I and the observed distributions are given in full in Table II.

The distributions shown in the first group are taken from the work of Rutherford and Geiger, Whitaker, Holm, and Greenwood and White. Rutherford and Geiger observed the collision with a small screen of an α particle emitted from a small bar of polonium placed at a short distance from the screen. The number of such collisions in each of 2608 eighth-minute intervals was recorded, the distance between bar and screen being gradually decreased so as to compensate for the decay of the radioactive substance. From this record two frequency-distributions were calculated, that of the number of particles striking the screen in an eighth-minute interval, and in a quarter-minute interval.[13] These are distributions (a1) and (a2), respectively. Distributions (a3) and (a4) are based on a count of the number of death notices in the London *Times* on each day for three consecutive years.[14] The distribution of deaths of men over 85 years of age (a3) and that of deaths of women over 80 (a4) are shown here. The next (a5) is a frequency-distribution of the number of telephone lines simultaneously in use, from measurements on a group of 100 subscribers.[15] The last distribution of this group (a6) was obtained from a count of the number of bacilli in each of 1,000 phagocytes, or white blood cells, in the same solution and as far as possible under the same conditions, and is typical of a large number of distributions of the number of tubercle bacilli ingested per cell.[16]

The first two examples in the second group are due to "Student" and the remaining seven are new. Distributions (b1) and (b2) show the results obtained from two different solutions of yeast cells by counting the number of cells per square of a haemacytometer slide on which the solution had been spread as uniformly as possible after it had been thoroughly shaken to break up any clumps of cells.[17] The next example (b3) was obtained from the records of the "lost and found" office of the Telephone and Telegraph Building, 195 Broadway, New York City. The number of lost articles found in the building

[13] "The Probability Variations in the Distribution of α Particles," by Ernest Rutherford and Hans Geiger, *Phil. Mag.*, Vol. 20, pp. 698–707, October, 1910.

[14] "On the Poisson Law of Small Numbers," by Lucy Whitaker, *Biometrika*, Vol. 10, pp. 36–71, 1914. Six other similar distributions are given.

[15] "Calculation of Blocking Factors of Automatic Exchanges," by Ragnar Holm, *P. O. E. E. J.*, Vol. 15, pp. 22–38, April, 1922.

[16] "A Biometric Study of Phagocytosis with Special Reference to the 'Opsonic Index'," by M. Greenwood and J. D. C. White, *Biometrika*, Vol. 6, pp. 376–401, 1908–1909. Fourteen other distributions are given.

[17] "On the Error of Counting with a Haemacytometer," by "Student," *Biometrika*, Vol. 5, pp. 351–360, 1906–1907. Two other distributions are given.



summary of the data used is given in Table I and the observed distributions are given in full in Table II.

The distributions shown in the first group are taken from the work of Rutherford and Geiger, Whitaker, Holm, and Greenwood and White. Rutherford and Geiger observed the collision with a small screen of an α particle emitted from a small bar of polonium placed at a short distance from the screen. The number of such collisions in each of 2608 eighth-minute intervals was recorded, the distance between bar and screen being gradually decreased so as to compensate for the decay of the radioactive substance. From this record two frequency-distributions were calculated, that of the number of particles striking the screen in an eighth-minute interval, and in a quarter-minute interval.[13] These are distributions (a1) and (a2), respectively. Distributions (a3) and (a4) are based on a count of the number of death notices in the London *Times* on each day for three consecutive years.[14] The distribution of deaths of men over 85 years of age (a3) and that of deaths of women over 80 (a4) are shown here. The next (a5) is a frequency-distribution of the number of telephone lines simultaneously in use, from measurements on a group of 100 subscribers.[15] The last distribution of this group (a6) was obtained from a count of the number of bacilli in each of 1,000 phagocytes, or white blood cells, in the same solution and as far as possible under the same conditions, and is typical of a large number of distributions of the number of tubercle bacilli ingested per cell.[16]

The first two examples in the second group are due to "Student" and the remaining seven are new. Distributions (b1) and (b2) show the results obtained from two different solutions of yeast cells by counting the number of cells per square of a haemacytometer slide on which the solution had been spread as uniformly as possible after it had been thoroughly shaken to break up any clumps of cells.[17] The next example (b3) was obtained from the records of the "lost and found" office of the Telephone and Telegraph Building, 195 Broadway, New York City. The number of lost articles found in the building

[13] "The Probability Variations in the Distribution of α Particles," by Ernest Rutherford and Hans Geiger, *Phil. Mag.*, Vol. 20, pp. 698–707, October, 1910.

[14] "On the Poisson Law of Small Numbers," by Lucy Whitaker, *Biometrika*, Vol. 10, pp. 36–71, 1914. Six other similar distributions are given.

[15] "Calculation of Blocking Factors of Automatic Exchanges," by Ragnar Holm, *P. O. E. E. J.*, Vol. 15, pp. 22–38, April, 1922.

[16] "A Biometric Study of Phagocytosis with Special Reference to the 'Opsonic Index'," by M. Greenwood and J. D. C. White, *Biometrika*, Vol. 6, pp. 376–401, 1908–1909. Fourteen other distributions are given.

[17] "On the Error of Counting with a Haemacytometer," by "Student," *Biometrika*, Vol. 5, pp. 351–360, 1906–1907. Two other distributions are given.

and turned in to the office on each day except Sundays and holidays was recorded and tabulated for the period from November 1, 1923 to September 30, 1925, inclusive, excluding June, July, and August of each year, when there might be considerable variations in the population of the building. Distribution (b4) shows the result of a count of the number of times that the number *12* appeared as the last two digits of a ten-place logarithm in a sample consisting of a column of 100 logarithms in Duffield's table,[18] and (b5) shows the number of fires per day in New York City in 1924, as reported daily in *The New York Times*, the figures for July 4 and for Election Day being discarded for obvious reasons. The last four examples in this group were taken from telephone company records of local service observations. A sample consisted of the calls observed at one central office in one month, and the series of samples used was selected from a complete record for all the central offices in a large city by the requirement that the number of calls per sample be not less than 450 nor more than 550. Distribution (b6) was obtained for the number of incorrect reports, (b7) for the number of cutoffs, (b8) for the number of double connections, and (b9) for the number of calls for the wrong number.

Group three is headed by Bortkewitsch's classical example of the Poisson exponential.[19] He found from the records of the Prussian army the number of men killed by the kick of a horse in each of 14 corps in each of 20 successive years, and, after discarding the records for 4 corps which were considerably larger than the others, treated the rest as one series of samples. This is distribution (c1). Series (c2) is similar to (b4), except that the samples of 100 two-place numbers were obtained from several different sources, logarithmic tables, trigonometric tables, and numbers listed in a telephone directory. Examples (c3), (c4), (c5), and (c6) show the variation in the number of telephone messages recorded per five-minute interval for certain groups of coin-box telephones in a large transportation terminal. The number of calls registered for each of 23 such telephones in each of about 20 five-minute intervals between noon and 2 p.m. was recorded on each of seven days (no Saturdays or Sundays included) but as the telephones are arranged in groups the distribution of the number of calls per interval was calculated for each group rather than for the individual telephones. These shown here are for a group of two telephones (c3), a group of four (c4), another group of two (c5), and a group of six (c6). The next four examples are

[18] "Logarithms, Their Nature, Computation, and Uses," by W. W. Duffield, Washington, 1897.

[19] Bortkewitsch, *op. cit.*

similar to examples (b6)–(b9), except that the limits of the number of calls per sample were 515 ± 25. Distribution (c7) was obtained for the number of cutoffs, (c8) for the number of double connections, (c9) for the number of calls for the wrong number, and (c10) for the number of connections to the wrong number. The next distribution (c11) was obtained from a count of the number of party-line subscribers listed per page of a large telephone directory and the last distribution of the group (c12) from a count of the number of advertisements in the "lost and found" column of *The New York Times* on each of the week-days from January 1, 1924 to August 31, 1924.

The fourth group contains only five examples, three of which are new. The first two of these present the same material used for example (b4) differently arranged. The 50,000 logarithms used are divided into 100 groups of 500 logarithms each for example (d1), and into 50 groups of 1,000 logarithms each for example (d2). The third (d3) is the distribution of the number of comets observed per year for the years 1789 to 1888 inclusive.[20] The other two distributions have been given by Perrin as typical of the data obtained when, in order to determine the density of the particles of an emulsion at a given depth, he restricted his field of vision to a tiny part of that layer, small enough so that the average number of particles visible was only one or two, and then made a large number of observations of the number of particles in that space at regular intervals.[21]

As was to be expected, these observed distributions have not only irregularities due to finite sampling but also in some cases what appear to be definite trends away from the corresponding Poisson distributions. In some cases there is an explanation ready at hand. For example, in series (b3), which gives the number of articles lost in the Telephone and Telegraph Building, the average number of articles lost per day might be expected to increase as the population of the building increased in this period following the completion of an addition, and the observed slant to the right is what would be expected. Also in series (d3), which gives the number of comets observed per year, the average would naturally increase steadily as a result of the continual improvement of telescopes and other instruments from 1789 to 1888. The curve toward the left in examples (c3) and (c5) might also be predicted because of the fact that the number of calls which could possibly be made in five minutes from a group of two telephones is certainly finite and probably rather small, and in examples (d4) and (d5) because it is difficult to judge by eye the number

[20] "Handbook of Astronomy," by G. F. Chambers, 4th ed., Oxford, 1889.

[21] "Brownian Movement and Molecular Reality," by Jean Perrin, London, 1910.

of particles visible simultaneously if that number is more than three or four.

In several cases special measures have been taken to reduce the variation of a and the resulting trend away from the corresponding Poisson distribution. In general, a is made as nearly constant as possible by making n and p constant throughout. In examples (b6)–(b9) and (c7)–(c10), for instance, each sample consists of approximately the same number of calls, and in example (c1) four corps were rejected because they were considerably larger than the others. In these examples it is assumed that p is practically constant and that by making n constant a constant average will be obtained. A somewhat different adjustment to keep a constant is illustrated by examples (a1) and (a2), where, as the decay of the radioactive substance decreases the average number of α particles emitted in a given solid angle per unit of time, the screen on which the particles strike is moved so that it intercepts a greater angle. In some cases n may be controlled much more easily than p, or vice versa, and a may be kept constant by letting one factor vary and adjusting the other to compensate, rather than by keeping both constant.

Summary

These examples of distributions which can be described by the Poisson exponential are of a dozen quite different kinds. They include eleven distributions found in published work on biometrics or statistics and twenty-one which are new. The agreement between the observed and the theoretical distribution is, in general, fairly good, and the applicability of the Poisson summation to a great variety of data is clearly indicated. The practical importance of some of these cases has been discussed above.

The use of the probability curves showing Poisson's exponential summation in place of double-entry tables as a source of data is shown to be simple, and their convenience as a background for plotting and comparing frequency-distributions is illustrated by Figs. 4 and 6–9. The new chart with a logarithmic scale for a (Fig. 5) is convenient in comparing distributions of different averages. It also shows the complete set of curves up to $a = 30$ instead of only to $a = 15$, and it makes it possible to read with considerable accuracy values of the variables in the range $0.1 \leqq a \leqq 2$, which is not clearly shown in Fig. 1 or Fig. 2.

TABLE II

c = number of occurrences of the event per sample.
m = number of samples with exactly c occurrences.
f = number of samples with at least c occurrences.
F = relative frequency of at least c occurrences per sample.

a1 *Alpha particles*
total = 10097
average = 3.87

c	m	f	F
0	57	2608	1.000
1	203	2551	.978
2	383	2348	.900
3	525	1955	.753
4	532	1440	.552
5	408	908	.348
6	273	500	.192
7	139	227	.087
8	45	88	.034
9	27	43	.0165
10	10	16	.0061
11	4	6	.0023
12	0	2	.00077
13	1	2	.00077
14	1	1	.00038

a2 *Alpha particles*
total = 10094
average = 7.74

c	m	f	F
0	0	1304	1.0000
1	3	1304	1.0000
2	17	1301	.9977
3	46	1284	.9847
4	99	1238	.949
5	126	1139	.873
6	151	1013	.777
7	187	862	.661
8	180	675	.518
9	173	495	.380
10	131	322	.247
11	75	191	.146
12	44	116	.089
13	35	72	.055
14	16	37	.028
15	14	21	.0161
16	1	7	.0054
17	1	6	.0046
18	2	5	.0038
19	1	3	.0023
20	1	2	.00153
21	1	1	.00077

a3 *Deaths of aged*
total = 903
average = 0.82

c	m	f	F
0	484	1096	1.000
1	391	612	.558
2	164	221	.202
3	45	57	.052
4	11	12	.0109
5	1	1	.00091

a4 *Deaths of aged*
total = 2364
average = 2.16

c	m	f	F
0	162	1096	1.000
1	267	934	.852
2	271	667	.609
3	185	396	.361
4	111	211	.193
5	61	100	.091
6	27	39	.036
7	8	12	.0109
8	3	4	.0036
9	1	1	.00091

a5 *Telephone lines in use**
total = ?
average = 4.32

c	M	F
0	.013	1.000
1	.045	.987
2	.125	.942
3	.185	.817
4	.187	.632
5	.186	.445
6	.126	.259
7	.071	.133
8	.036	.062
9	.018	.026
10	.005	.008
11	.002	.003
12	.001	.001

a6 *Bacilli*
total = 1927
average = 1.93

c	m	f	F
0	219	1000	1.000
1	267	781	.781
2	219	514	.514
3	129	295	.295
4	70	166	.166
5	50	96	.096
6	26	46	.046
7	13	20	.020
8	5	7	.007
9	2	2	.002

b1 *Yeast cells*
total = 720
average = 1.80

c	m	f	F
0	75	400	1.000
1	103	325	.813
2	121	222	.555
3	54	101	.253
4	30	47	.118
5	13	17	.043
6	2	4	.0100
7	1	2	.0050
8	0	1	.0025
9	1	1	.0025

b2 *Yeast cells*
total = 1872
average = 4.68

c	m	f	F
0	0	400	1.000
1	20	400	1.000
2	43	380	.950
3	53	337	.843
4	86	284	.710
5	70	198	.495
6	54	128	.320
7	37	74	.185
8	18	37	.093
9	10	19	.048
10	5	9	.023
11	2	4	.010
12	2	2	.005

b3 *Lost articles*
total = 439
average = 1.04

c	*m*	*f*	*F*
0	169	423	1.000
1	134	254	.600
2	74	120	.284
3	32	46	.109
4	11	14	.033
5	2	3	.0071
6	0	1	.0024
7	1	1	.0024

b4 *Number 12*
total = 421
average = 0.84

c	*m*	*f*	*F*
0	231	500	1.000
1	150	269	.538
2	92	119	.238
3	24	27	.054
4	1	3	.006
5	1	2	.004
6	1	1	.002

b5 *Fires***
total = 9487
average = 26.1

c	*m*	*f*	*F*
0		364	1.0000
5		364	1.0000
10		363	.9973
15		346	.951
20		286	.786
25		185	.508
30		103	.283
35		53	.146
40		22	.060
45		18	.049
50		8	.022
55		4	.0110

b6 *Incorrect reports*
total = 138
average = 0.27

c	*m*	*f*	*F*
0	388	506	1.000
1	102	118	.233
2	12	16	.032
3	4	4	.0079

b7 *Cutoffs*
total = 1057
average = 2.09

c	*m*	*f*	*F*
0	75	506	1.000
1	126	431	.852
2	141	305	.603
3	73	164	.324
4	50	91	.180
5	29	41	.081
6	6	12	.024
7	2	6	.0119
8	3	4	.0079
9	0	1	.0020
10	0	1	.0020
11	1	1	.0020

b8 *Double connections*
total = 1760
average = 3.48

c	*m*	*f*	*F*
0	21	506	1.000
1	63	485	.958
2	98	422	.834
3	97	324	.640
4	85	227	.449
5	61	142	.281
6	42	81	.160
7	18	39	.077
8	11	21	.042
9	6	10	.0198
10	3	4	.0079
11	0	1	.0020
12	0	1	.0020
13	0	1	.0020
14	0	1	.0020
15	1	1	.0020

b9 *Calls for wrong number*
total = 2520
average = 4.98

c	*m*	*f*	*F*
0	10	506	1.000
1	20	496	.980
2	45	476	.941
3	60	431	.852
4	85	371	.733
5	92	286	.565
6	73	194	.383
7	55	121	.239
8	28	66	.130
9	18	38	.075
10	9	20	.040
11	5	11	.022
12	3	6	.0119
13	2	3	.0059
14	1	1	.0020

c1 *Deaths from kick of horse*
total = 122
average = 0.61

c	*m*	*f*	*F*
0	109	200	1.000
1	65	91	.455
2	22	26	.130
3	3	4	.020
4	1	1	.005

c2 *Number 12*
total = 251
average = 1.00

c	*m*	*f*	*F*
0	90	250	1.000
1	95	160	.640
2	46	65	.260
3	15	19	.076
4	3	4	.016
5	0	1	.004
6	0	1	.004
7	1	1	.004

c3 *Calls from group of two coin-box telephones*
total = 172
average = 1.19

c	*m*	*f*	*F*
0	44	145	1.000
1	48	101	.697
2	38	53	.366
3	13	15	.103
4	1	2	.0138
5	1	1	.0069

c4 *Calls from group of four coin-box telephones*
total = 384
average = 2.74

c	m	f	F
0	5	140	1.000
1	33	135	.964
2	24	102	.729
3	38	78	.557
4	23	40	.286
5	9	17	.121
6	4	8	.057
7	4	4	.029

c5 *Calls from group of two coin-box telephones*
total = 212
average = 1.50

c	m	f	F
0	27	141	1.000
1	49	114	.809
2	39	65	.461
3	19	26	.184
4	7	7	.050

c6 *Calls from group of six coin-box telephones*
total = 468
average = 3.39

c	m	f	F
0	8	138	1.000
1	13	130	.942
2	20	117	.848
3	37	97	.703
4	24	60	.435
5	20	36	.261
6	8	16	.116
7	5	8	.058
8	2	3	.022
9	1	1	.0072

c7 *Cutoffs*
total = 557
average = 2.09

c	m	f	F
0	44	267	1.000
1	62	223	.835
2	71	161	.603
3	43	90	.337
4	25	47	.176
5	14	22	.082
6	4	8	.030
7	2	4	.0150
8	2	2	.0075

c8 *Double connections*
total = 906
average = 3.39

c	m	f	F
0	14	267	1.000
1	33	253	.948
2	48	220	.824
3	56	172	.644
4	43	116	.434
5	34	73	.273
6	22	39	.146
7	8	17	.064
8	4	9	.034
9	3	5	.0187
10	2	2	.0075

c9 *Calls for wrong number*
total = 1351
average = 5.06

c	m	f	F
0	3	267	1.000
1	12	264	.989
2	23	252	.944
3	31	229	.858
4	45	198	.742
5	50	153	.573
6	37	103	.386
7	29	66	.247
8	13	37	.139
9	12	24	.090
10	4	12	.045
11	4	8	.030
12	3	4	.0150
13	1	1	.0037

c10 *Connections to wrong number*
total = 2334
average = 8.74

c	m	f	F
2	1	267	1.0000
3	5	266	.9963
4	11	261	.978
5	14	250	.936
6	22	236	.884
7	43	214	.801
8	31	171	.640
9	40	140	.524
10	35	100	.375
11	20	65	.243
12	18	45	.169
13	12	27	.101
14	7	15	.056
15	6	8	.030
16	2	2	.0075

c11 *Party lines*
total = 1981
average = 6.60

c	m	f	F
0	7	300	1.000
1	9	293	.977
2	14	284	.947
3	17	270	.900
4	21	253	.843
5	40	232	.773
6	46	192	.640
7	42	146	.487
8	32	104	.347
9	17	72	.240
10	22	55	.183
11	12	33	.110
12	6	21	.070
13	10	15	.050
14	1	5	.0167
15	3	4	.0133
16	0	1	.0033
17	1	1	.0033

c12 *"Lost and found" advertisements***
total = 7051
average = 33.7

c	m	f	F
0		209	1.0000
5		209	1.0000
10		208	.9952
15		207	.9904
20		199	.952
25		182	.871
30		144	.689
35		93	.445
40		51	.244
45		21	.100
50		7	.033
55		2	.0096

d1 *Number 12*
total = 421
average = 4.21

c	m	f	F
0	2	100	1.00
1	6	98	.98
2	18	92	.92
3	13	74	.74
4	16	61	.61
5	19	45	.45
6	13	26	.26
7	5	13	.13
8	5	8	.08
9	2	3	.03
10	1	1	.01

d2 *Number 12*
total = 421
average = 8.42

c	*m*	*f*	*F*
3	1	50	1.00
4	5	49	.98
5	2	44	.88
6	6	42	.84
7	6	36	.72
8	5	30	.60
9	7	25	.50
10	6	18	.36
11	4	12	.24
12	5	8	.16
13	1	3	.06
14	0	2	.04
15	2	2	.04

d3 *Comets*
total = 258
average = 2.58

c	*m*	*f*	*F*
0	19	100	1.00
1	19	81	.81
2	17	62	.62
3	14	45	.45
4	13	31	.31
5	8	18	.18
6	4	10	.10
7	2	6	.06
8	3	4	.04
9	1	1	.01

d4 *Particles in emulsion*
total = 46
average = 0.92

c	*m*	*f*	*F*
0	22	50	1.00
1	16	28	.56
2	7	12	.24
3	4	5	.10
4	1	1	.02

d5 *Particles in emulsion*
total = 106
average = 2.12

c	*m*	*f*	*F*
0	6	50	1.00
1	11	44	.88
2	12	33	.66
3	14	21	.42
4	6	7	.14
5	1	1	.02

* M is the relative frequency of exactly c occurrences per sample. Holm does not state the actual number of samples from which this was calculated, but it was evidently at least 1000.

** Since in the range $a > 30$ the curves are drawn only for every fifth value of c, in these two distributions which extend beyond $a = 30$ the values of f and F are tabulated only for every fifth value of c, and the values of m, which are meaningless unless the complete series is given, are omitted.

Line Current Regulation in Bridge Polar Duplex Telegraph Circuits

By S. D. WILBURN

Synopsis: A mathematical analysis of the bridge polar duplex telegraph circuit, under the condition that the bridge arms are of equal resistance, shows that there is a particular bridge arm resistance which results in maximum received current. As the bridge arm resistances are increased beyond the value giving this maximum, the received current diminishes gradually. On the other hand, as the bridge arm resistances are decreased below the value giving the maximum, the received current drops off very rapidly. It follows that when necessary to limit line current, the maximum received current is obtained by placing the regulating resistance in the bridge arms. Also when the line resistance is large enough to limit the line current to less than the maximum allowable value, a gain may be obtained by increasing the bridge arm resistance to the value which corresponds to maximum received current. Experience has shown that in many situations where difficulty is encountered in operating a duplex telegraph circuit with the regulating resistances *in the line*, a very decided improvement is obtained by transferring these resistances to the bridge arms.

FOR the operation of polar duplex telegraph circuits, line batteries of uniform voltage are generally used and it is usually desirable to maintain the line current within fairly definite limits. The most suitable line battery voltage and the desired limits for the line current depend upon the type of line and apparatus used. In order to maintain the line current within the desired limits with uniform voltage it is necessary to add resistance to the circuit in greater or less amounts depending upon the length and gauge of the line circuit used. On account of line trouble and the necessity for rerouting telegraph circuits for other reasons, it is frequently desirable to switch a duplex set from one line to another of different resistance. To facilitate line current regulation without delaying service when such changes in line assignment are made, it is of considerable operating advantage to include in the wiring of each duplex circuit an adjustable resistance in the form of a rheostat mounted in an accessible location at the duplex set so that the attendant can readily regulate the line current at the time that necessary adjustments in the balancing artificial line are made to suit the changed line condition.

This paper outlines an investigation which was made with the object of finding an arrangement of line current regulating resistance which would result in the maximum steady-state received current with the bridge duplex telegraph circuit shown by Fig. 1, where it is desired to limit the line current to about .070 ampere. The condition for maximum steady-state received current was sought as the first step toward determining the most suitable arrangement of the resistances with the viewpoint that such an arrangement would probably be the most

satisfactory from a transmission standpoint if it did not adversely effect the important factor of received current wave shape. An arrangement of the resistances was found which results in the maximum steady-state received current and from oscillographic tests which were subsequently made, this arrangement fortunately appears to improve the wave shape of the received current as compared with that resulting from other possible arrangements considered. It was also found from field trials on a number of practical circuits that this arrangement

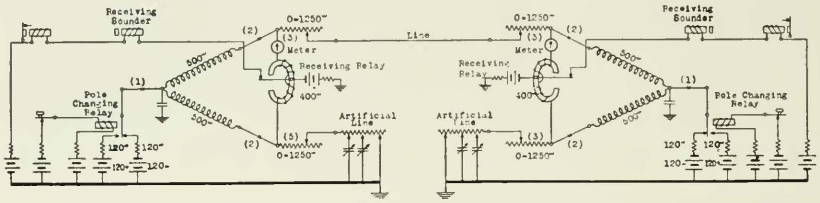


Fig. 1—Bridge Duplex Telegraph Circuit

results in improved transmission over other arrangements which have been considered for the regulating resistances.

Three different locations in the bridge duplex circuit are considered for the regulating resistances. These locations are designated (1), (2) and (3) in Fig. 1 and may be described respectively as follows:

(1) A single resistance in series with the battery branch of the circuit.

(2) Equal resistances in series with each of the bridge arms.

(3) Equal resistances in series with the line and the artificial line of the duplex set.

In considering locations (2) and (3), it is assumed that the resistances are in the form of a double rheostat with the movable arms mechanically connected to facilitate adding equal amounts of resistance simultaneously.

It will be seen from the circuit shown by Fig. 1 that of the three locations for the regulating resistance, (3) might be expected to reduce the received current most for a given line current, as that arrangement introduces resistance directly between the receiving relays. However, as that location for the resistances had been in general use, and since it was not at all obvious which of the other two arrangements would be the most favorable from the standpoint of received current, it seemed desirable to set up line current and received current equations to determine how the currents would be affected by the resistances in each

location. Of the six current equations required, the one for the received current with the resistance in location (2) was found to possess a maximum within a resistance range which made that arrangement the most favorable from the standpoint of steady-state received current.

Curves i_1, i_2 and i_3 Fig. 2, show the steady-state value of received current which will be obtained on lines of 500 to 2260 ohms resistance

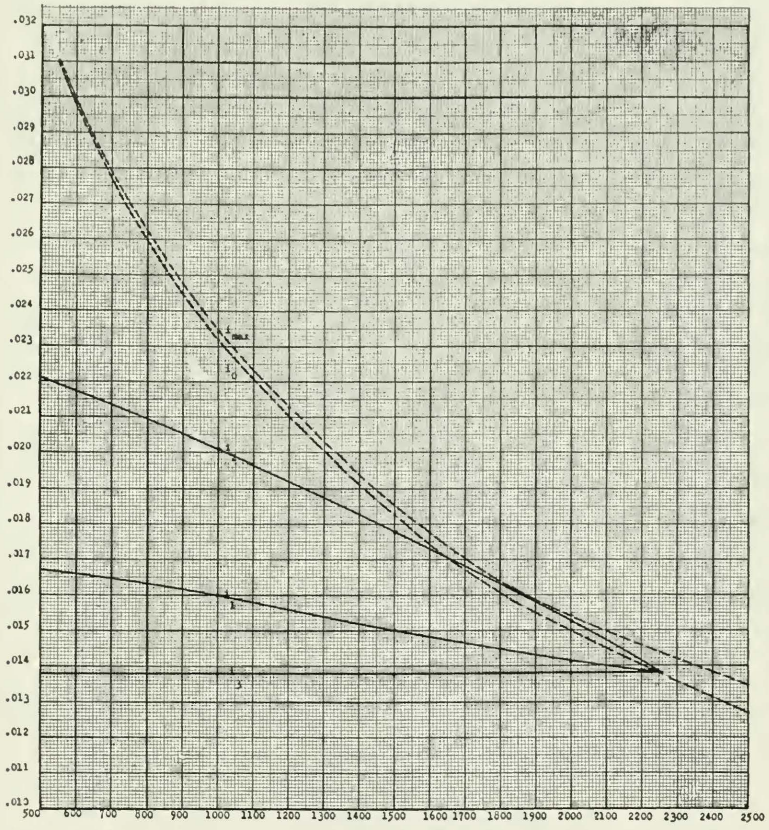


Fig. 2

with the regulating resistances located at points (1), (2) and (3), respectively. In each case just sufficient resistance is added to make the line current .070 ampere. If the resistance of the line is greater than 2260 ohms, the line current will fall below .070 ampere without the addition of resistance at either point. It will later be shown that, regardless of line current limitations, location (2) results in the maximum

practicable steady-state received current in bridge duplex operation the bridge arms being of equal resistance.[1]

The method of calculating curves i_1, i_2, and i_3 Fig. 2, will be discussed presently along with certain other mathematical considerations.

In setting up the equations for the received current and line current, certain practical operating conditions of the circuit are assumed; first, that the circuit as a whole be kept symmetrical by using the same amount of regulating resistance at each station and second, that the duplex sets be maintained in a state of balance for direct currents. Line leakage will be neglected.

To express the line and received currents as direct, or explicit functions of the regulating resistances under the assumed condition of the circuit requires the use of unusually cumbersome equations which may to some extent be avoided in the early part of the solution without sacrificing accuracy. The complicated nature of the equations is due largely to the intricate relation between the regulating resistances and the overall network resistance of the duplex set from the terminal of the line to ground and, in turn, the relation between this network resistance and the two currents. With the exception of one step in the present investigation the work has been shortened by representing this network resistance by a parameter, or second independent variable, r which is itself a quadratic function of the regulating resistance, represented by R. The required values of r are then computed from the equation connecting it to R. In the expressions for the ratios of received current to line current all determinants which cannot be readily reduced to the second order cancel out so that considerable work is avoided by using these ratios rather than the explicit line current equations for calculating the line currents.

The equations expressing the relation between the received currents, i_1, i_2 and i_3 and the regulating resistance in the three locations (1), (2) and (3) respectively, are as follows:

$$i_1 = \frac{aE}{(T+R_1+r_1)\,(2a+b)+ab+a^2} \tag{1}$$

$$i_2 = \frac{R_2E}{(T+G+r_2)\,(2R_2+b)+bR_2+R_2^2} \tag{2}$$

$$i_3 = \frac{aE}{(T+R_3+G+r_3)\,(2a+b)+ab+a^2} \tag{3}$$

[1] For the case of unequal bridge arms see Heaviside's "Electrical Papers," Vol. I, p. 24.

and the expressions for the ratios of the received currents to corresponding line currents, I_1, I_2 and I_3 are:

$$\frac{i_1}{I_1} = \frac{a(1/2T+r_1)}{(T+r_1)(2a+b)+ab} \tag{4}$$

$$\frac{i_2}{I_2} = \frac{R_2(1/2T+r_2)}{(T+r_2)(2R_2+b)+R_2b} \tag{5}$$

$$\frac{i_3}{I_3} = \frac{a(1/2T+r_3)}{(R_3+T+r_3)(2a+b)+ab} \tag{6}$$

where,

a, represents the constant resistance of each bridge arm in arrangements (1) and (3);

b, the resistance of the receiving relay;

E, the voltage of the line battery which is assumed to be equal at both stations and may be either negative or positive;

G, the constant resistance in series with the line battery taps in arrangements (2) and (3);

T, the resistance of the line between the duplex sets.

R_1, R_2 and R_3 are the regulating resistances in the different locations corresponding to the subscripts. In the equations for arrangement (1), G is assumed to be contained in R_1 and in the equations for arrangement (2), a is assumed to be contained in R_2. The equations for the parameters, r_1, r_2 and r_3 are as follows:

$$r_1 = \sqrt{\tfrac{1}{4}T^2+R_1T+\frac{aT(a+b)+ab(a+2R_1)}{2a+b}} - \tfrac{1}{2}T \tag{7}$$

$$r_2 = \sqrt{\tfrac{1}{4}T^2+GT+\frac{R_2T(R_2+b)+bR_2(R_2+2G)}{2R_2+b}} - \tfrac{1}{2}T \tag{8}$$

$$r_3 = \sqrt{\tfrac{1}{4}T^2+GT+R_3(T+R_3+2G)+\frac{2aR_3(a+b)+aT(a+b)+ab(a+2G)}{2a+b}} - \tfrac{1}{2}T \tag{9}$$

While the line current and the received current can be calculated for any values of R and T from equations (1) to (9) inclusive, explicit line current equations are needed for calculating the received current for a definite value of line current, such as shown by curves i_1, i_2 and i_3, Fig. 2. It is clear that these curves cannot be calculated from equa-

tions (1) and (9) alone, as the first step necessary is to determine the value of R which, with a given value of T, will result in the specified value of I (.070 ampere). With line current equations R can, of course, be calculated by substituting .070 for I. While the line current equations can be set up fairly readily, they are of an extremely cumbersome character. For that reason curves i_1, i_2 and i_3, Fig. 2, were calculated by the following method:

From equations (1) to (9) inclusive, the line current was calculated for the various values of T from 500 to 3000 ohms with various values of R from 0 to 2000 ohms in steps of 250 ohms. For each value of T the line current was then plotted against R and the required value of the latter read from the intersection of the curve and the .070 ordinate. The values of R thus obtained were then substituted in equations (7) to (9) for calculating r. These values of R and r in turn were substituted in equations (1) to (3). By the above method the values of R within plus or minus two or three ohms can be determined. This possible error in R will not appreciably effect the points on the curves. The point of intersection of i_1, i_2, i_3, and i_0, Fig. 2 was calculated by equating the right hand side of equation (10) to .0138.

Referring to equations (1), (2) and (3) showing the relations between the regulating resistances and the received currents, it will be noted that in the right hand member of (1) and (3), R_1 and R_3, respectively, appear only as positive terms in the denominator. This shows that the received current will inevitably be reduced for every increase in the resistance, provided r_1 and r_3 are continuously increasing functions of R_1 and R_3 and from equations (7) and (9) it will be seen that both r_1 and r_3 increase continuously for every increase in R_1 and R_3, respectively. In equation (2), however, R_2 appears in both the numerator and the denominator and in the latter it appears in both the first and second powers. It is, therefore, not so easy to determine from an inspection of the equation just how the received current will be affected by increasing the resistance. It will be seen that this difference in the received current equations offers a guide in the selection of the location for the resistances which will result in the greatest received current.

From a closer inspection of equation (2), it is seen that when $R_2 = 0$ the received current will be 0 and, as the denominator of the right hand member contains the second power of R_2, the received current will approach 0 if R_2 be increased indefinitely. Also, it is clear that there will be current in the receiving relay for all finite values of R_2. Thus, if R_2 be indefinitely increased from 0, the received current will rise from 0 to a maximum value and then descend again toward 0.

This suggests solving for the value of R_2 corresponding to the point where i_2 is a maximum by differentiating equation (2) with respect to R_2 and equating to 0. The nature of the equation shows also that i_2 will have but one maximum. If the value of R_2 corresponding to maximum i_2 proves to be greater than 500 ohms, it will open up the possibility of increasing the received current by adding the regulating resistances at points (2), Fig. 1.

In calculating the line and received currents for different values of R_2 it is, of course, permissible to calculate separately corresponding values of r_2 and then substitute these values as constants in equation (2). Obviously this procedure cannot be followed in finding the derivative of i_2 with respect to R_2. The expression to be dealt with in this differentiation is that which results from the substitution of the right hand member of equation (8) for r_2 in equation (2). This substitution gives the following explicit and rigorous equation for the steady-state current in the receiving relays of a balanced symmetrical bridge duplex telegraph circuit:

$$i_2=\frac{ER_2}{(\frac{1}{2}T+G)(2R_2+b)+R_2(R_2+b)+\sqrt{T(\frac{1}{4}T+G)(2R_2+b)^2+(2R_2+b)[R_2T(R_2+b)+bR_2(R_2+2G)]}} \tag{10}$$

Equation (10) was found useful in calculating received currents as it combines (2) and (8) and may be used instead of equations (1) and (7) by changing G to R_1 and R_2 to a, but when it is differentiated and equated to 0 the resulting equation for R_2 corresponding to maximum received current is of an extremely impractical nature as it involves various powers of R_2 up to the sixth, together with an unusually large number of terms. In this investigation, it was not necessary to solve this equation for R_2 as it was found that for values of R_2 and T within the practical ranges of 500 to 1750 ohms for R_2 and 500 to 3000 ohms for T, r_2 is very nearly equal to $1/3\ R_2+2\sqrt{T}+200$. If this expression be substituted for r_2 in equation (2) and the result differentiated and equated to 0 it leads to the following equation which gives values of R_2 corresponding fairly close to the point of maximum received current:

$$R_2=\sqrt{\tfrac{3}{4}b(T+2T^{\frac{1}{2}}+G+200)} \tag{11}$$

With a receiving relay of 400 ohms resistance and a battery tap resistance of 120 ohms, as shown in Fig. 1, equation (11) becomes

$$R_2=10\sqrt{3(T+2T^{\frac{1}{2}}+320)}$$

From this equation, it is found that the bridge arms, each consisting of a 500 ohm bridge coil only, as shown by Fig. 1, are too small for maximum received current if the line resistance is greater than approximately 555 ohms. With line circuits ranging in resistance from 1000 to 2500 ohms, the respective values of R_2 necessary for maximum received current strength, range from approximately 624 to 935 ohms. If, then, resistance be added in the proper amounts at the points designated (2), Fig. 1, the received current will be increased thereby and at the same time, the line current will be reduced. If the line circuit resistance is approximately 1650 ohms or more the amount of resistance needed at points (2) to make the received current maximum, will be sufficient to reduce the line current to .070 ampere or less. This is illustrated by the three upper curves in Fig. 2. The lower broken curve, designated i_0, represents the received current which will be obtained with no regulating resistance in the circuit at either point. It will be seen that this curve passes below curve i_2 at a point corresponding to a line resistance of 1650 ohms. With approximately that value of line resistance and no regulating resistance, the line current is approximately .086 ampere and the received current is .0171 ampere. If approximately 410 ohms be added at points (2) the line current will be reduced to .070 ampere and the received current will remain at .0168 ampere. Curve $i_{\text{max.}}$ shows the maximum received current which can be realized by adding correct amounts of resistance at points (2). The upper curve touches curve i_2 at a point corresponding to a line resistance of 1850 ohms. That is, with a line resistance of this value, the regulating resistance required to reduce the line current to .070 ampere is just sufficient to bring the received current up to the maximum. For lines of this resistance or greater, the line current can be reduced to .070 ampere or less and at the same time the received current is increased. It will be seen from Fig. 2, that as compared to locations (1) and (3) for the regulating resistance, the advantage of location (2) from a steady-state received current standpoint, becomes greater with lines of low resistance and amounts to 32.3% and 60.1% respectively, with a line of 500 ohms resistance. On the other hand, the increase in received current due to arrangement (2), as compared to the condition of no regulating resistance, becomes greater with lines of higher resistance, as shown by the divergence of the i_2 and $i_{\text{max.}}$ curves, Fig. 2.

With line resistances in the lower range, the amount of regulating resistance needed to make the received current maximum will not be enough to bring the line current down to the desired value of .070 ampere. For example, with a line of 500 ohms resistance, the 500 ohm

bridge arms are already too large by approximately 14 ohms and 1470 ohms will be required at points (2) to bring the line current down to .070. The bridge arms will then be 1484 ohms greater than needed for maximum received current. The question then arises as to why arrangement (2) results in the subtsantial received current gains with lines of low resistance, as shown by curves i_1, i_2 and i_3, Fig. 2. This part of the problem can best be solved by plotting equation (10). Fig. 3 shows this equation plotted for a 1200 ohm line. It will be seen that, from the

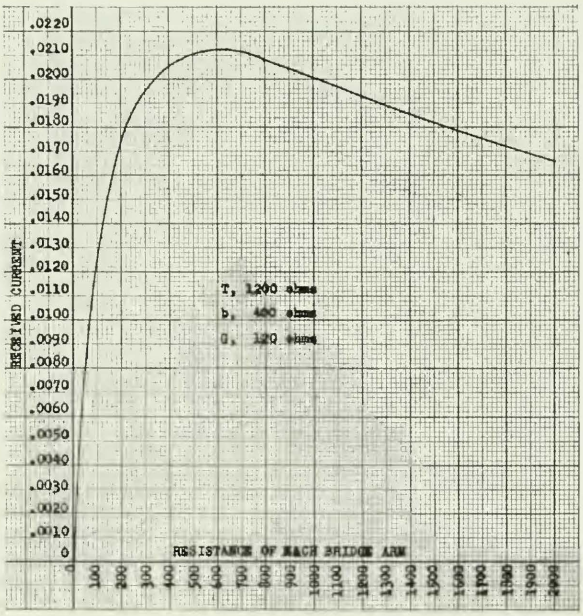


Fig. 3

standpoint of received current strength, it is better to have the bridge arm too great than too small, as the received current rises rapidly to a maximum and then descends slowly. On the other hand, if resistances be added at points (1) or (3), the operating point on the received current curve will in all cases be moved further away from the maximum, and this movement away from the maximum will take place on the side of the maximum which has the greatest effect in reducing the received current, as will be shown.

The resistance at points (1) or (3) moves the operating point on the received current curve away from the maximum due to the fact that the value of the bridge arm resistance corresponding to maximum

received current is a function of both the resistance between the duplex sets and the resistance in the battery branch of the circuit, as shown by equation (11); G in this equation corresponds to R_1 in equation (1) and T represents all the resistance between the duplex sets, this being augmented by the addition of resistance at points (3). Therefore, any increase in the resistance in the battery branch of the circuit or in the resistance between the sets, such as will result

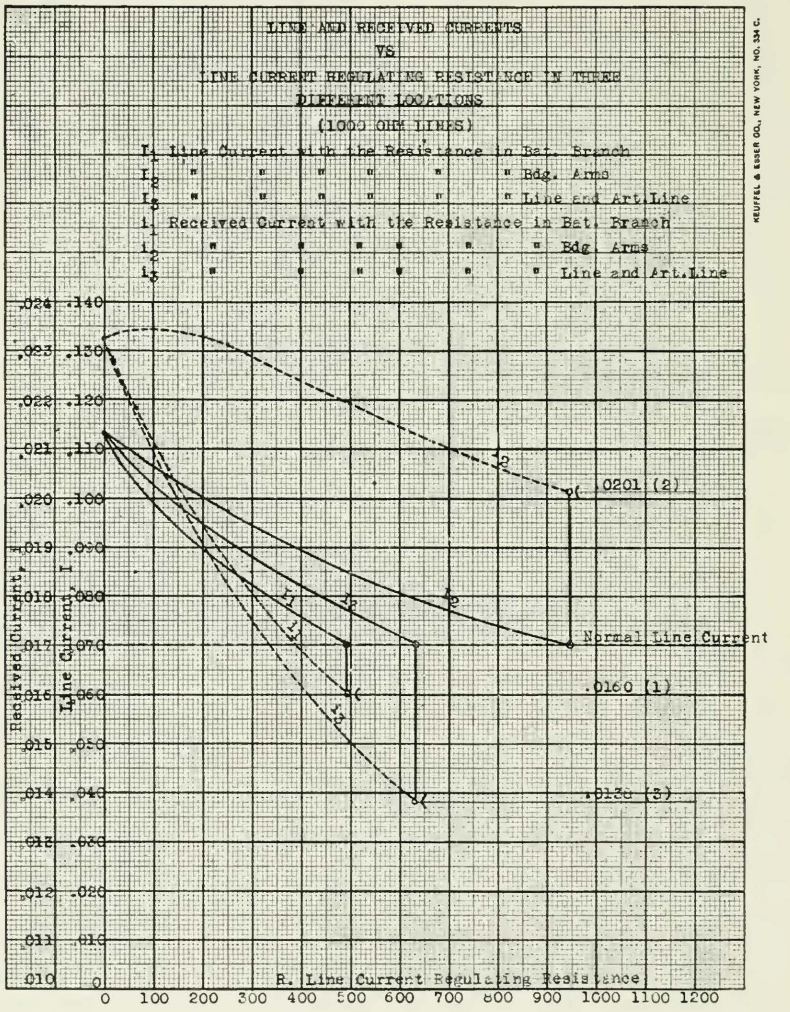


Fig. 4—Line and Received Currents vs Line Current Regulating Resistance in Three Different Locations—(1000 Ohm Line)

from adding resistance at points (1) or (3) respectively, will move the maximum point on the curve further to the right. This can best be illustrated by the following example:

With a line of 1500 ohms resistance, the resistance required in each bridge arm for maximum received current is about 750 ohms, so that the normal 500 ohm bridge arms as shown by Fig. 1 are short of the maximum by 250 ohms. If the line current be reduced to the desired value of .070 ampere by adding resistance at points (3) about 250 ohms will be required at each station. This will make the resistance between the duplex sets, corresponding to T in equation (11), $1500+500=2000$, for which the value of the bridge arms for maximum received current is about 855 ohms. The operating point on the curve is, therefore, 355 ohms on the left hand side of the maximum, as compared with 250 ohms before the resistances were added. The change in the maximum due to adding resistance at points (1) takes place in the same general way though not in exactly the same degree.

Fig. 4 shows how the line and received current are affected by the resistances in each location with a 1,000 ohm line. From these curves it will be noted that location (2) for the resistances results in a gain of about 25.6 per cent. in received current strength as against location (1) and as compared to location (3) the gain in received current amounts to about 45.6 per cent.

As the ratio of the bridge arms is not changed by adding the line current regulating resistance in equal amounts at points (2) that arrangement should introduce no difficulties in maintaining a balance between the line and artificial line. Furthermore, arrangement (2) should not increase disturbances due to small extraneous currents in the line.

Carrier-Current Communication on Submarine Cables

Los Angeles-Catalina Island Telephone Circuits[1]

By H. W. HITCHCOCK

SYNOPSIS: Seven telephone channels and one telegraph channel on one single-conductor deep-sea cable have been made possible by the employment of carrier current on one of the two submarine cables across Catalina channel. This is the only application of carrier telephony to deep-sea cables and the system is one of the shortest carrier systems (26 mi.) in commercial operation; it provides more separate carrier channels (six) than has been previously attempted; and it differs in other important respects from other systems. This paper describes this carrier-current system.

IN the commercial application of new developments in the electrical communication art, there are a few places which repeatedly call attention to themselves. Notable among these is Catalina Island, for it is probable that in providing telephone service across the short expanse of water which separates Catalina from the mainland, more novel improvements have been employed than at almost any other point.

The first commercial telephone communication with Catalina Island was established in 1920 when a radio system was placed in operation between Avalon and the mainland, the circuit being extended by wire to Los Angeles. This circuit was in use for several years and featured in a number of transcontinental demonstrations, including the one which was held at the opening of the service to Havana over the Key West-Havana cables.

The system is of considerable interest as it represents the only instance in which radio has been used, in this country at least, to form a portion of a toll telephone system for the general use of the public. That it was reasonably successful is demonstrated by the fact that on some days as many as 183 commercial telephone messages and a large number of telegrams were handled over it. The system also proved to be one of the first popular broadcasting stations and many letters were received from radio fans, often several hundred miles away, telling of some of the amusing conversations which were overheard.

In 1923 the radio was replaced by two single-conductor submarine cables. By that time the demands for service were too great to be met by a single circuit, while the growing interest in radio broadcasting, as well as the increasing interference from ship transmitters,

[1] Presented at the Pacific Coast Convention of the A. I. E. E., Salt Lake City Utah, Sept. 6–9, 1926.

rendered its continued operation very difficult and unsatisfactory. The submarine cables were of the single-conductor, deep-sea type, each providing a single-wire circuit. They are of interest for a number of reasons, chiefly, perhaps, because they represent one of the few instances of deep-sea cable manufacture in this country. From the cable hut at San Pedro, the circuit is extended to the office by means of a special lead-covered cable containing four individually shielded No. 13 B & S gauge pairs for the telephone circuits and four 19-gauge pairs for the telegraph circuits and other miscellaneous uses. Between the San Pedro office and Los Angeles, the circuit was composed of a No. 19 B & S gauge cable phantom. At San Pedro a through-line repeater was inserted in order to secure the desired over-all equivalent between Avalon and Los Angeles.[2]

Although the two circuits provided by the cables represented a great improvement over the previous condition as regards the quality of the service rendered and the number of messages which could be handled, it was realized that they would soon prove inadequate to handle the heavy summer business, for which eight or ten circuits would be required in a relatively short time. To provide for such a large increase by the laying of additional cables was deemed impracticable, as the cost would be excessive. Furthermore, in water of this depth—3,000 feet—it is important that cables be laid at least a mile or two apart, so that in the event that trouble develops on one, it can be repaired without disturbing any of the others. For a total distance as short as the width of the Catalina channel—23 nautical miles—such a separation between adjacent cables could not be maintained without materially increasing the length of the outer ones with a corresponding increase in their cost and in their transmission equivalents. In view of these facts, it was decided to secure as many more circuits as possible by operating carrier systems over the two cables already in use. This project was actively promoted with the result that on May 15, 1926, six carrier telephone circuits were placed in operation.

The use of carrier in the past few years has increased so rapidly that the mere addition of a new system is, in itself, of hardly more than passing interest. In this instance, however, there are a number of factors which render the project of particular interest. It is one of the shortest carrier systems—26 miles—in commercial operation. It is the only application of carrier telephone to deep-sea cables; the system pro-

[2] A description of these cables and their laying was given in a paper presented by the writer at the Pacific Coast Convention in 1923 and published in Volume XLII of the *Transactions*.

vides more separate channels (six) than has ever before been attempted, while the particular arrangement employed is different in many other important respects from anything which has been used in the past.

In order to better appreciate the reasons for adopting the system finally agreed upon, it may be of interest to review briefly the essential characteristics of carrier systems and the different types which are available.[3]

Carrier systems may be divided into two general classes, namely, balanced or grouped, depending upon the manner in which the currents in the two directions are prevented from interfering with each other

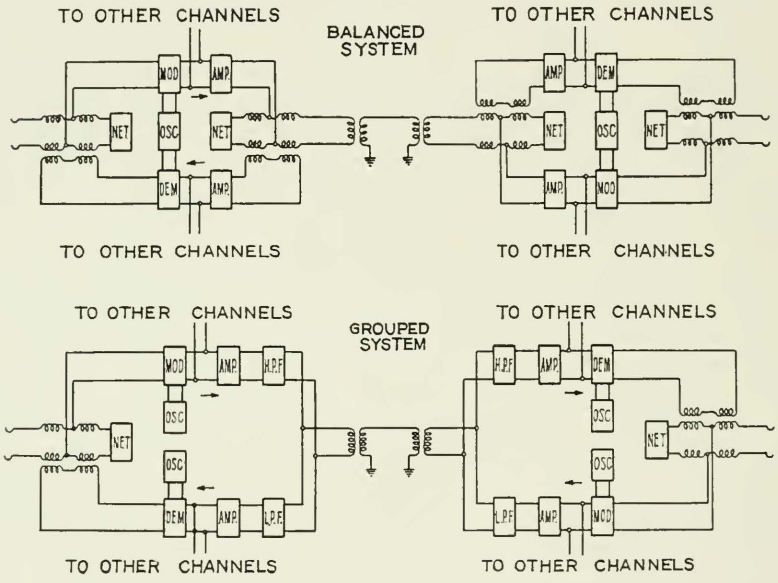


Fig. 1

at the terminals. In the balanced system this separation is accomplished by means of a three-winding transformer or hybrid coil together with a balancing artificial line such as is used with a voice-frequency repeater. In the grouped system, different carrier frequencies are used for transmission in the two directions and their separation at the terminals is effected by means of suitable band-pass filters. These two systems are shown diagrammatically in Fig. 1. The balanced system has the advantage that for each channel

[3] The general principles of carrier-current telephony are described at considerable length in a paper by Messrs. Colpitts and Blackwell which was published in Volume XL of the Journal of the Institute.

the same carrier frequency may be used for transmission in both directions so that there may be as many channels as there are separate carrier frequencies. On the other hand, the wire circuit must be very uniform throughout so that the impedance will be very regular over the entire carrier-frequency range, and may be simulated by an artificial line. The line must also be very stable so that the impedance balance, once having been secured, will not be disturbed. Furthermore, as transmission with the same carrier takes place in the two directions, the effect of the cross-talk between systems of the same type is very severe, so that it is usually impracticable to operate two of these over wires which are in close proximity for any considerable distance. The grouped system has the advantage that a balancing line is not required and hence small circuit irregularities are relatively unimportant. Furthermore, the effect of cross-talk is much less severe, so that a number of systems may often be operated over adjacent circuits. One disadvantage is that two carrier frequencies are required for each channel so that fewer circuits can be secured with one system.

Carrier systems may also be divided into two classes depending upon the manner in which the carrier current is provided at the receiving end. In the carrier transmission system, the carrier current is supplied by the oscillator at the sending end and is transmitted over the circuit along with one or both of the side bands. In the carrier suppression system, the carrier current itself is not transmitted but is introduced into the receiving equipment from a local source. This latter system is proving to be superior for general carrier purposes because of the advantages which accrue from relieving the line and apparatus from the load of the carrier current.

Turning now to the electrical characteristics of the cables, we find that each one provides a circuit having a transmission equivalent which increases throughout the carrier range but is moderate in magnitude. The impedance, as is to be expected with a uniform, non-loaded cable, is very smooth, and since there is no opportunity for any change in the cable constants, the impedance has practically no variation. The transmission equivalent and the impedance of one of the cables are shown in Figs. 2 and 3, respectively. The cross-talk between the cables is small enough to be entirely negligible, regardless of the type of carrier systems employed.

In view of all the conditions outlined, a balanced system of the carrier suppression type was decided upon. Such a system provides the maximum number of channels per cable, while the usual difficulties of impedance balance and inter-system cross-talk are largely

absent due to the unusual characteristics of the cables. The adoption of such a system also made possible the employment of standard units of equipment of the most recent design. The general nature of the system and the arrangement of the component parts is shown diagrammatically in Fig. 4. Fig. 5 is a simplified circuit diagram

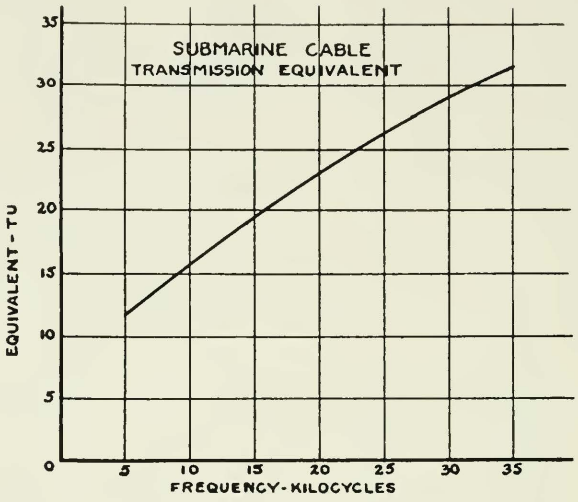


Fig. 2

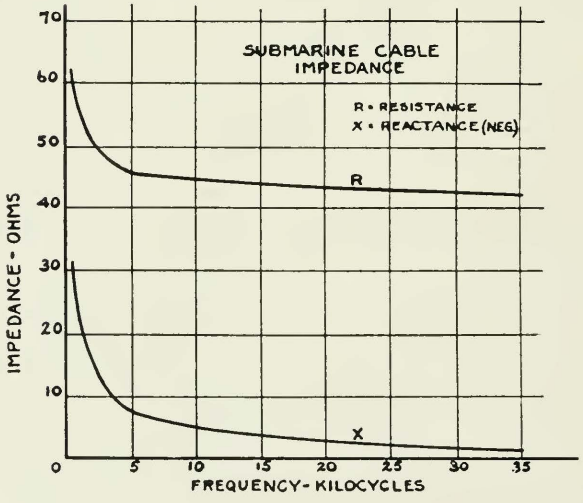


Fig. 3

showing the filters for separating the various circuits at the terminals, together with the balancing arrangement. In Fig. 6 are shown the essential parts of one channel together with the amplifiers and the hybrid coil which are common to all the channels. For convenience, some of the battery and auxiliary circuits have been omitted in the figure.

At the time the system was under development, it was uncertain that balanced operation of all channels over a single cable would be practicable, so that an alternative arrangement involving substan-

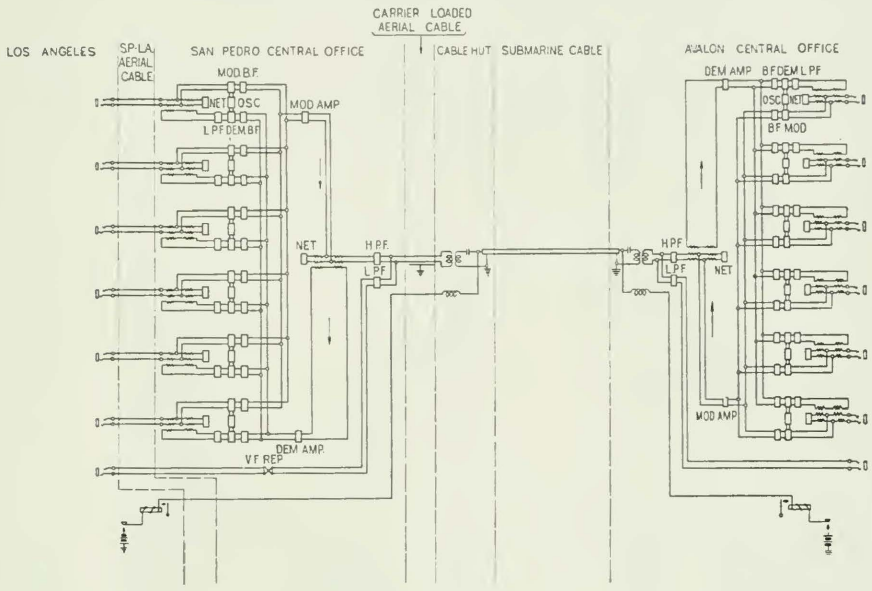


Fig. 4

tially four-wire operation over the two cables was provided for. With this arrangement, which is shown diagrammatically in Fig. 7, all transmission in one direction takes place over one cable, while transmission in the opposite direction is effected over the second cable. No balancing equipment or hybrid coils are employed. Such an arrangement would increase the system stability, if such were required, but would limit the total carrier capacity of the two cables to six channels. In the event of the failure of one cable, operation with such a system would be impossible, and it would be necessary, at that time, to revert to the two-wire arrangement as described above, with a possible reduction in the over-all gain or a reduction in the number of operating channels.

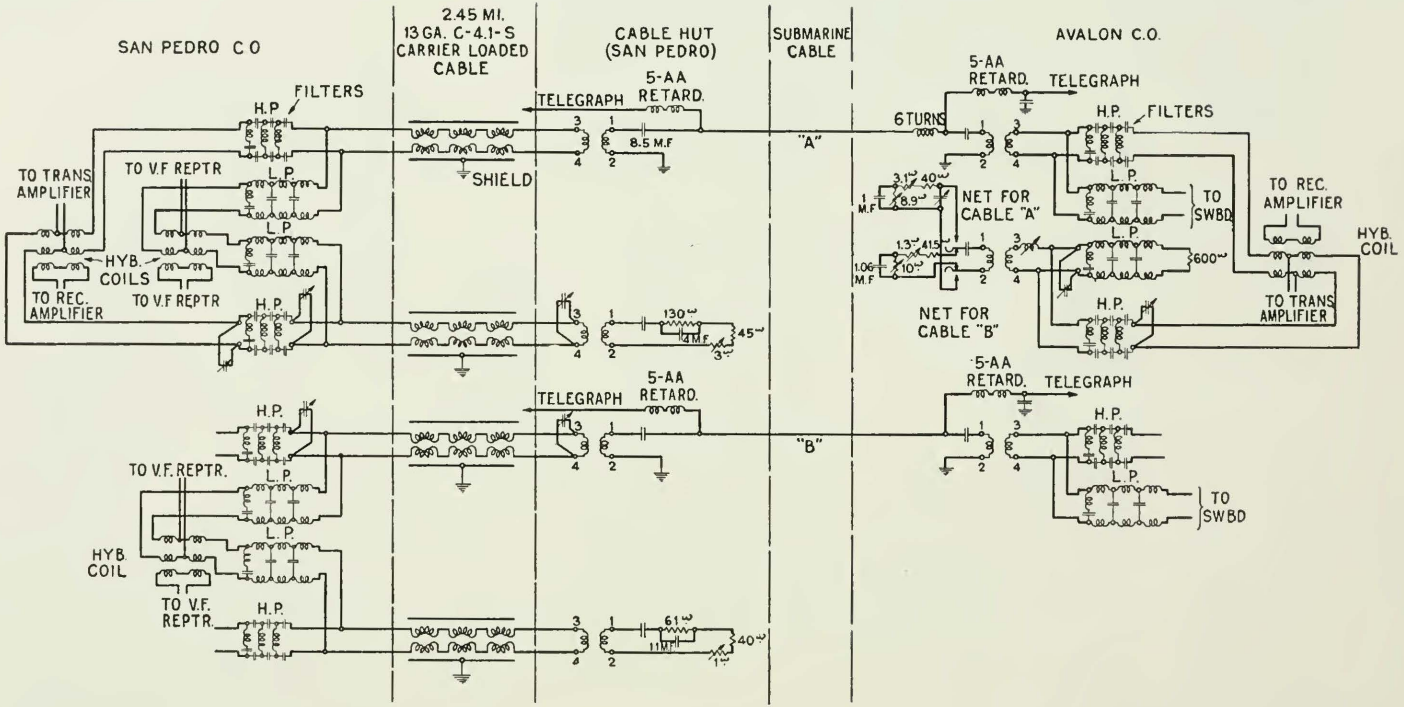

SAN PEDRO C O
2.45 MI. 13 GA. C-4.1-S CARRIER LOADED CABLE
CABLE HUT (SAN PEDRO)
SUBMARINE CABLE
AVALON C.O.
FILTERS
H.P
L.P
TO TRANS AMPLIFIER
TO V.F REPTR
HYB. COILS
TO REC. AMPLIFIER
TO V.F REPTR
TELEGRAPH
5-AA RETARD.
8.5 M.F
SHIELD
"A"
6 TURNS
NET FOR CABLE "A"
TO SWBD
TO REC. AMPLIFIER
HYB. COIL
600Ω
NET FOR CABLE "B"
TO TRANS AMPLIFIER
130Ω
4 M.F
45Ω
"B"
TO V.F. REPTR.
HYB. COIL
TO V.F. REPTR.
61Ω
11 M.F
40Ω

Fig. 5

As may be seen from Fig. 4, a carrier-equipped cable provides a d-c. telegraph circuit, and one voice-frequency and six carrier-frequency telephone channels. The separation of the various channels is effected by means of electrical filters. Fig. 8 shows the band of frequencies employed for each channel. For the d-c. telegraph this

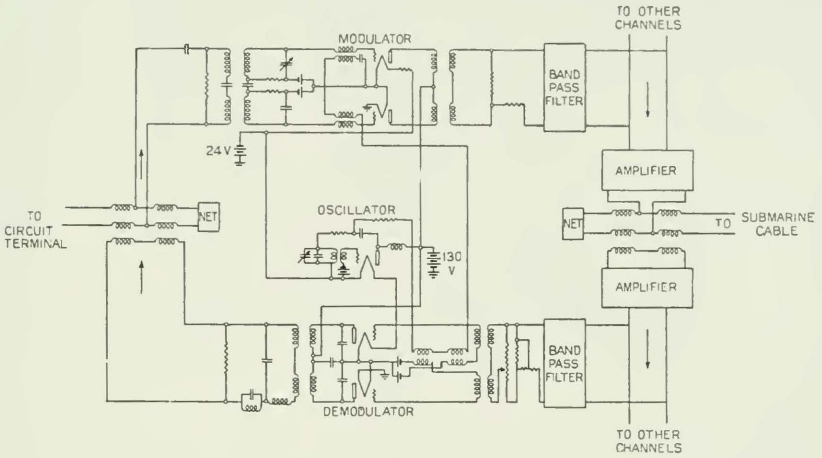


Fig. 6

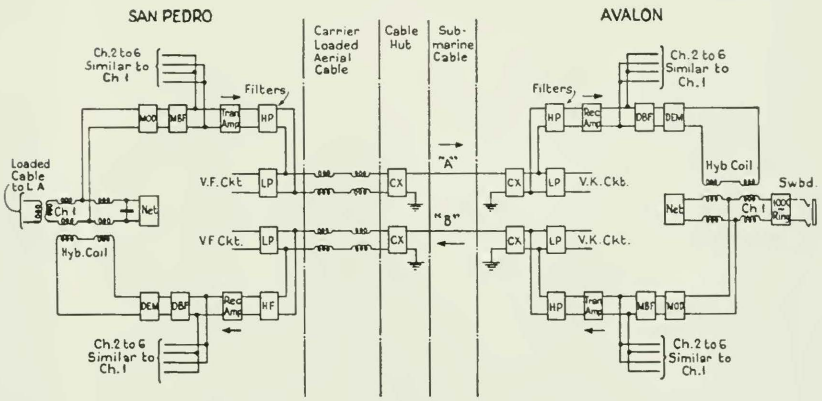


Fig. 7

separation is effected at the terminals of the cable as is shown in Fig. 5. The telegraph circuit requires a continuous d-c. path, whereas the telephone channels require the insertion of an inequality ratio insulating transformer at the ends of the cable in order to properly join the 43-ohm grounded cable circuit with the 600-ohm metallic

circuit formed by the office equipment and intermediate cable. As this transformer must pass both the voice and carrier channels, it has been designed so as to have a high efficiency for all frequencies between 250 and 30,000 cycles. Separation of the voice-frequency circuit from the carrier system is performed by means of the usual high and low pass filters which are located at the central offices. These filters both have a cut-off frequency of 3,000 cycles, the low

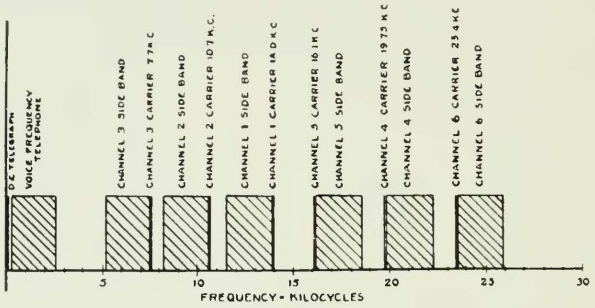


Fig. 8

pass transmitting all frequencies below this value and the high pass transmitting all above it. In the carrier system the transmitting and receiving currents are separated from each other by a hybrid coil and balancing network. Between the output of the six modulators and the common transmitting amplifier, individual band-pass filters are located. Each one of these filters is designed to transmit one of the side bands produced by the modulator associated with it and to suppress all other frequencies. The six receiving currents are separated in a similar manner. Each filter allows current of the proper frequency to pass to the corresponding demodulator and excludes all others. Each demodulator is also provided with a low pass filter which allows the passage of the resulting voice-frequency current but excludes all incidental higher frequencies which might be present and render the circuit noisy. The input of each modulator and the output of the corresponding demodulator are finally joined by means of a voice-frequency hybrid coil and extended to the circuit terminal as a two-wire circuit. At the San Pedro end, each two-wire circuit is extended to Los Angeles over a loaded cable circuit. Phantoms are employed for this purpose as they have a higher cut-off frequency than have the side circuits, with a correspondingly better quality.

Concerning the carrier system itself, the two ends are practically identical while the general equipment arrangement for an individual

channel is the same in all cases except for the frequency of the band-pass filter. For this reason, a consideration of one channel is sufficient. Each channel is composed of a voice-frequency hybrid coil, a modulator with its band-pass filter, an oscillator, and a demodulator, together with its associated filters. In addition, there is, at each end, a carrier hybrid coil together with transmitting and receiving amplifiers which are common to all channels. The arrangement of this equipment is shown schematically in Fig. 6, as previously indicated.

The modulator, the input of which is connected to the center taps of the hybrid coil line windings, utilizes of two vacuum tubes arranged for push-pull operation. The carrier current which is supplied by the oscillator is applied to the two grids by means of a transformer. Such a circuit generates the two side bands but suppresses the carrier. In order that this suppression may be as complete as possible, the small condenser associated with the grid of one of the tubes is made variable and is adjusted until the carrier current in the modulator output is reduced to a minimum. The band-pass filter transmits one of the side bands and suppresses the other, as well as all miscellaneous resultant currents of a higher order which are produced by the modulator. It also prevents the output currents of the other channels from entering the modulator circuit as this would cause a reduction in their efficiency and give rise to undesirable frequencies.

The demodulator is very similar to the modulator. The tube arrangement is substantially the same and carrier current is supplied from the one oscillator. In the demodulator a complete suppression of the carrier is unnecessary as this is accomplished by the low pass output filter. For this reason, the small balancing grid condensers are omitted. In order to adjust the over-all gain of the channel, the demodulator is provided with an adjustable potentiometer graduated in two transmission unit steps, and in addition, fixed pads are provided for making further gain adjustments. The output of the demodulator is connected to the series winding of the voice-frequency hybrid coil.

The oscillator which supplies the carrier current to the modulator and demodulator is of the usual type. The tuning condenser includes a small variable unit for making small adjustments in frequency. Separate oscillators are used at the two ends for each channel, and as these are in no way connected together, it is occasionally necessary to make slight adjustments in order to keep the frequencies at the two ends substantially equal. The oscillators are very stable, however, and such adjustments are seldom required.

The individual channel filters are all of the band-pass type as previously indicated and have a free transmission range of approximately 2,500 cycles. Outside this free range they have a high impedance so as not to act as a shunt for the other channels. They are all of substantially the same construction, although the constants of the component parts necessarily vary as the filters for the different channels transmit different frequencies.

The transmitting and receiving amplifiers, which are practically identical, are shown schematically in Fig. 9. They consist of two push-pull stages connected in tandem. Each half of the second or output stage consists of two parallel tubes of high output capacity. In this way a comparatively high gain and a large energy output

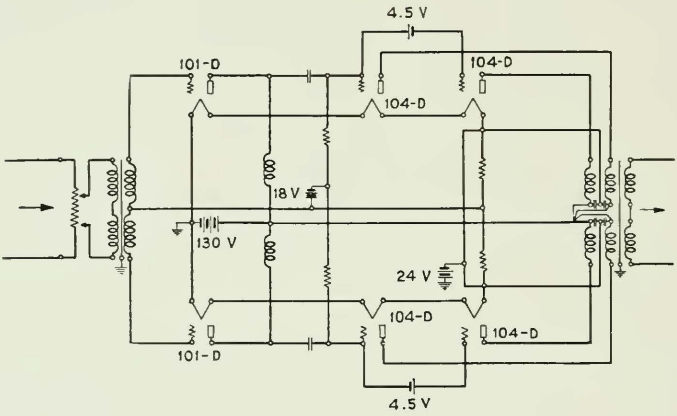


Fig. 9

may be secured without overloading. This is very important as these amplifiers are common to all six channels and any tendency to overload would produce objectionable distortion and inter-channel modulation. In order to adjust the over-all gain for the entire system, each amplifier is provided with an input potentiometer.

As has been previously indicated, the transmitting and receiving circuits are joined to the cable by means of hybrid coils. Probably the most difficult problem encountered in the installation of this system was the securing of an adequate balance. The difficulty of doing this may be better appreciated when it is realized that this balance must cover all frequencies from 3,000 to 30,000 cycles, and must have a value of from 30 to 45 T. U., the higher value which represents an impedance unbalance of approximately one per cent. being required at the upper frequency. In order to secure such a

balance, every part of the line circuit was matched by a similar part in the network circuit. All filters and transformers on the line side of the hybrid coil were duplicated in the network, and on the San Pedro side a 13-gauge carrier-loaded cable pair was included in the network circuit between the office and the cable hut, and the inequality ratio transformer and basic network simulating the cable were located at the latter point. In addition to providing a balance within the carrier range, it was necessary at the San Pedro end for the network circuit to balance the cable within the voice-frequency range as a through-line repeater is employed on the voice-frequency circuit. Not only was it necessary to duplicate all parts in the line

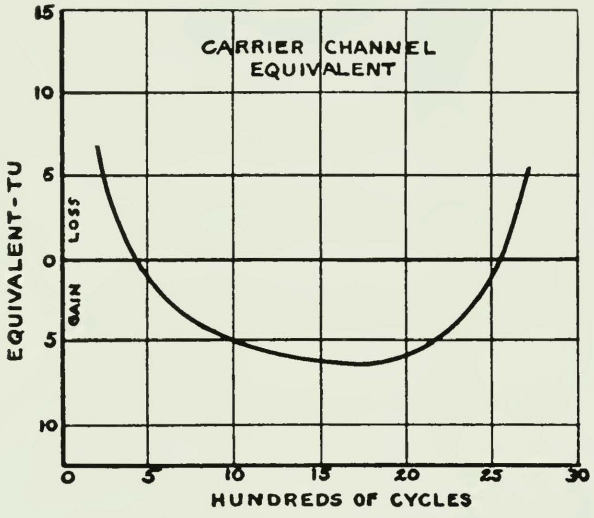


Fig. 10

and network circuits but in addition they were carefully selected and paired so that the two parts associated would have, as nearly as possible, the same electrical characteristics. All wire pairs within the office which appeared in the carrier frequency circuits were individually shielded by means of a grounded metallic covering. The 13-gauge carrier-loaded pairs in the cable joining the hut and the office were also individually shielded by means of a lead foil wrapping. This was done in order to preserve the balance and prevent crosstalk with another system which may be placed on the second cable at some future time.

Although extreme care was exercised in making the refinements described, the balance was still lower than was desired so that small

variable auxiliary impedances were inserted at suitably chosen points in the line and network circuits. By the adjustment of these elements, it was found that the balance could be raised to any desired value for any particular channel, but that in so doing, the balance on some of the others would be impaired. By careful adjustment, however, it was possible to secure a balance for all channels within the range previously mentioned. As the transmission equivalent of the cable

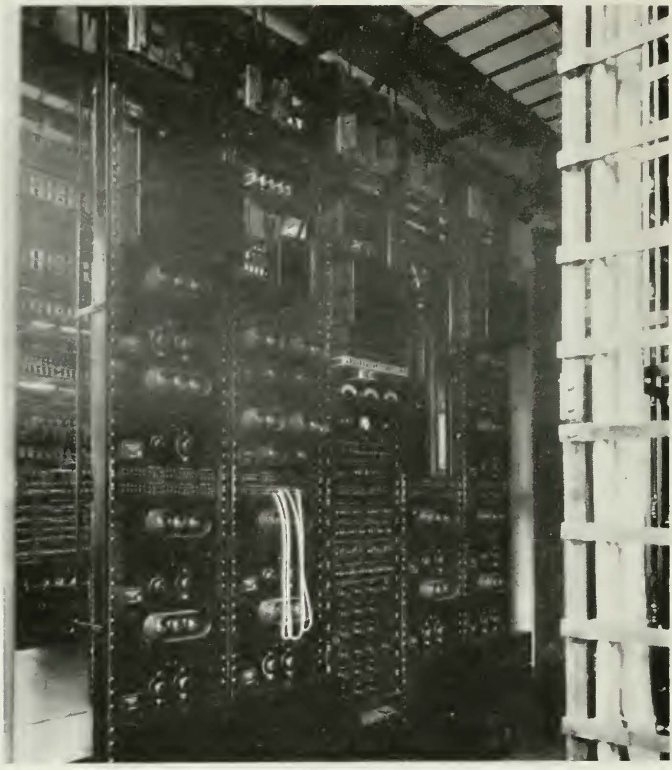


Fig. 11

increases with the frequency, the over-all channel gains must be increased in the same manner in order that all circuits may have the same over-all equivalent. The networks were therefore arranged so that the higher frequencies would have the better balance, as in that way the margin of balance over gain could be made substantially the same for all channels. Since this margin should not be allowed to fall below a fairly definite minimum if the circuit is to have the

desired stability, it is evident that the balance which may be secured determines the over-all gain which is possible. In this case the circuit equivalent for all channels between Los Angeles and Avalon was set at five T. U. As the loaded cable between Los Angeles and San Pedro is approximately nine T. U., it may be seen that the carrier system actually introduces a gain and performs the function of a

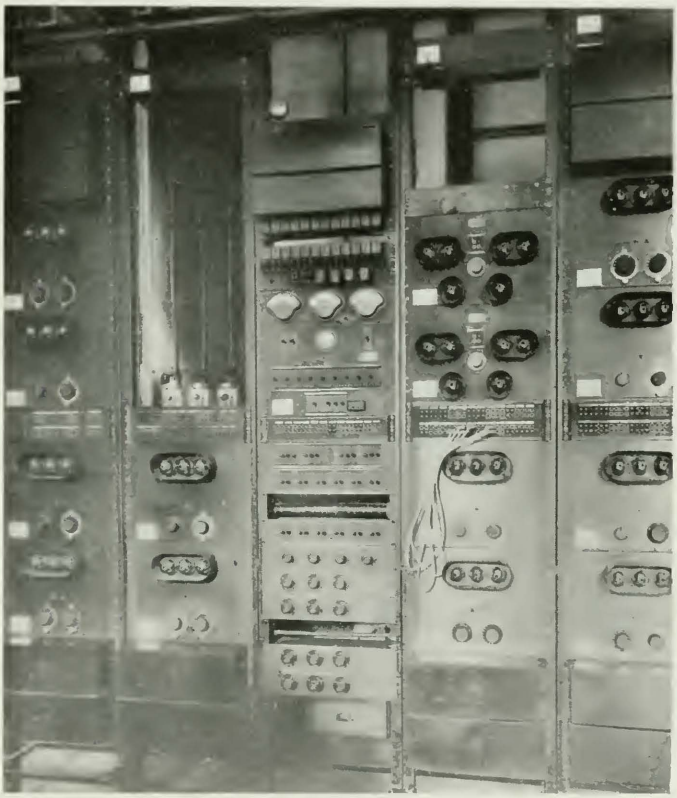


Fig. 12

repeater besides increasing the number of circuits. Fig. 10 gives a frequency characteristic of one of the channels which is typical of all of them. Balancing equipment has been provided for both cables as is shown in Fig. 5. With this arrangement, the carrier system may be operated over either cable. The transfer from one cable to the other is so simple that it can be made with practically no traffic interruption.

Signaling over the carrier channels is effected by means of 1,000-cycle ringers which are connected to the circuits at the two terminals. As the ringing current is within the voice range, it is transmitted over the regular carrier channel so that no additional signaling equipment is necessary.

In order to insure satisfactory operation, all necessary testing facilities are included. Meters and keys are provided for measuring the voltages of the plate, grid and filament batteries as well as the

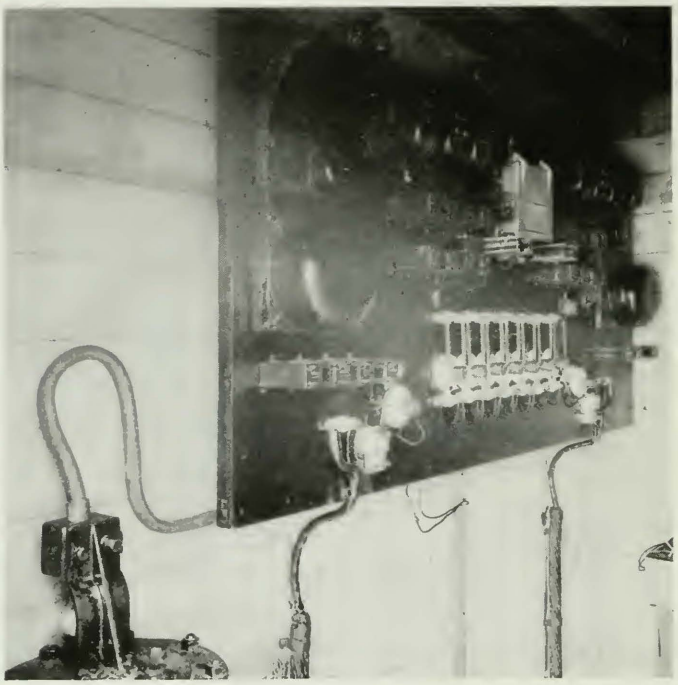


Fig. 13

plate and filament currents of all tubes. Individual rheostats are inserted in all filament circuits for making any adjustments that may be necessary. Alarms are provided to indicate any abnormal condition which might develop on any tube. Thermocouples and artificial lines have been conveniently arranged for checking the efficiency of all units such as the modulators and demodulators. Jacks are located at suitable points so that any changes which may be necessary can be quickly made.

The general appearance of the carrier system may be seen from Figs. 11 and 12 which show the equipment at San Pedro and Avalon respectively. Fig. 13 is an interior view of the San Pedro cable hut showing the cable terminals, together with the insulating transformers, telegraph composite sets, and basis networks. Referring to the central office equipment, the first bay contains the equipment for two complete channels. At the top are the terminal strips for making all connections with the equipment below. On the next two small panels are mounted the hybrid coils and the other miscellaneous apparatus associated with the voice-frequency ends of the two channels. Below these are the modulator and demodulator band filters which are covered with dust proof cases. Next comes the modulator and demodulator panels for one channel. Below the two jack strips is mounted similar equipment for a second channel but arranged in reverse order. In the upper half of the second bay is located a small panel mounting the carrier hybrid coil and associated equipment. Below this appear the transmitting and receiving amplifiers. The lower half of the bay is similar to the lower half of the first one. In the third bay is mounted all the battery supply and testing apparatus. The first two units contain the battery retard coils. Below these are the alarm relays and auxiliary resistances. Next come the meters for measuring the tube currents and voltages, and below these are the thermocouples and artificial lines for making high frequency measurements. Below the jack strip are the keys for opening and closing the individual filament circuits used for measuring the plate and filament currents. Alarm lamps are also associated with each of the filament circuits. The fourth bay is similar to the second except that the upper half is vacant. As may be seen from the photographs, the amplifiers appear on the second bay at San Pedro and on the fourth at Avalon. The fifth bay is an exact duplicate of the first.

The new system has now been in successful operation for the past five months. In the light of its performance thus far, we feel assured, that when more circuits are required a second system of six channels can be added to the second cable, thus providing a total of fourteen telephones and two telegraph circuits over the two single-conductor cables. Such a circuit group, we believe, will meet the traffic requirements for quite a number of years.

Abstracts of Recent Technical Books and Papers from Bell System Sources

Theory of Vibrating Systems and Sound. I. B. CRANDALL. Recent years have witnessed a great revival in Acoustics, both as a subject of industrial research and as a field for academic cultivation. This development has been carried on most actively in America and in Germany, and has been measurably due to the demands of the public for commercial acoustic devices in the public address and radio fields. In addition, there has evolved the new subject of Architectural Acoustics, largely through the researches of the late W. C. Sabine and his followers in this country.

In this situation, with an ever increasing body of technical literature, it may be noted that the standard textbooks on Sound have maintained their classical character, although new impedance methods and terminology have come into use (closely parallel to those of electrical theory) and many new fields of study have been opened up. The need for a connected treatment involving the new subject matter and methods the present author has attempted to supply, and while necessarily building on the classical treatises as a foundation, he has included a chapter on complex vibrating systems with a treatment of acoustic filters; two chapters on the theories of sound generation and radiation with applications to tubes and horns and a chapter on the essentials of architectural acoustics and absorbing materials. An extended bibliographic appendix serves as an entry to many branches of Applied Acoustics—for example, loud speaking telephones, piezoelectric resonators, recent work on speech and hearing, and submarine signalling, to cite only a few examples.

The author has purposely avoided duplicating classical material such as the theories of bars or of Fourier's series, feeling that the newer ideas and developments deserve the emphasis gained thereby.

The book aims to organize Acoustics as now practiced, and research workers will probably find it useful, not only in their classes, but as a starting point for acoustic research.

The book is produced by D. Van Nostrand & Co. ($5.00).

Contemporary Physics. KARL K. DARROW. During the twenty-five years since this century began, the science of physics has undergone amazing enlargements and transformations which may well be ranked among the most significant attainments of our times. Through

discoveries and improvements in the arts of experimentation, it became possible to measure the charge and the mass of the atom of negative electricity; to measure the charges, masses and magnetic moments of the atoms of the chemical elements; to study the processes of detaching atoms of electricity from atoms of matter; and to extend the spectra of the elements by detecting a host of radiations previously unknown and determining their frequencies. The data so assembled, together with observations upon the encounters of electrified particles with atoms, illuminated the relations between the elements, and contributed to the design of an atom-model which has already inspired many discoveries. Among these the greatest was the discovery of the Stationary States, which replaced the early way of interpreting spectra by a new and strikingly fruitful procedure, and taught experimenters to seek after and find a multitude of new phenomena of the most varied interest and importance in almost every field of physics. To name only two of the fields thus enriched: the flow of electricity through gases, and the conditions for the excitation of radiation, have been clarified in most unexpected fashion since the recognition of the Stationary States.

The book "Contemporary Physics," by Dr. Darrow, is devoted to these fundamental discoveries and to some of their consequences. It might be described as an introduction to the Theory of Atomic Structure, in the present day acceptance of this phrase; for the phenomena with which it deals have nearly all been used in designing atom models, and reciprocally a great many of them have been discovered in the course of making experimental tests of predictions based upon atom-models. These models are in fact among the most important features of contemporary physics and it would neither be possible nor desirable to omit them from such a book; for they are undoubtedly valuable, and the phenomena could hardly be described briefly and clearly without making use of them. Nevertheless, the actual facts of experience receive the greater emphasis, for they are the permanent and unassailable parts of the recent extensions of physics.

The book is developed from articles which have been appearing in the BELL SYSTEM TECHNICAL JOURNAL under the general title *Some Contemporary Advances in Physics*, apart from the articles concerning Hertzian waves and conduction of electricity in solids, which fall outside of the field to which it is restricted. The remaining articles, which were originally self-contained and separate, were largely rewritten in order to build out of them a coherent presentation of a unified field, and in the course of the rewriting they were nearly doubled in length by addition of new material. The book may

be described as an elementary treatise, and although hardly of popular nature it will be intelligible to anyone who has had a fairly thorough college course in physics.

Published by D. Van Nostrand & Co. ($6.00).

Electric Circuit Theory and the Operational Calculus. JOHN R. CARSON. This book is based upon a course of fifteen lectures delivered recently at the Moore School of Electrical Engineering of the University of Pennsylvania.

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. The present author ventures the suggestion that 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 his hope that this method of approach and exposition will secure 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 volume 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, the author has not hesitated to employ developments and explanations 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 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 upon whether the reader's viewpoint is that of the mathematician or the engineer.

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 distinctly pedagogic. To electrical engineers and to electrical instructors this exposition of the fundamentals of electric circuit theory and the operational calculus should be of more than ordinary value. An appendix furnishes a list of original papers and memoirs which gives a fairly complete bibliographic survey of the field.

The volume is published by McGraw-Hill Book Co., price $3.00.

Exploring Life: the Autobiography of Thomas A. Watson. To have been the youth who at the age of twenty was assigned to build Alexander Graham Bell's original telephone apparatus, and then to share with him and Sanders and Hubbard the cares of rearing the telephone industry in the United States to healthy childhood, and finally to share the handsome profits which accrued therefrom, would doubtless satisfy the desires of the average ambitious individual. But to look upon this as merely a beginning and before the age of thirty to separate himself voluntarily from the business he had helped to found and set forth in quest of achievement in other and entirely unrelated fields attests an eagerness to play the game of life which cannot fail to be an inspiration to everyone.

So interesting is the life the author has led and so charmingly has he related his varied activities that the book would be welcome at any time, but coming during the semicentennial of the invention of the telephone, it is appropriate as well. To those who are desirous of obtaining further light on the early career of the telephone, particularly in the United States, the book brings several chapters of new material. But to the much wider circle who find enjoyment in a document which is at once homely and adventurous, every page of this autobiography will yield delight.

Published by D. Appleton & Co., price $3.50.

Some Measurements of Short Wave Transmission.[1] R. A. Heising and J. C. Schelleng and G. C. Southworth. Quantitative data on field strength and telephonic intelligibility are given for radio transmission at frequencies between 2.7 megacycles (111 m.) and 18 megacycles (16 m.) and for distances up to 1,000 miles, with some data at 3,400 miles. The data are presented in the form of curves and surfaces, the variables being time of day, frequency and distance. Comparisons are made between transmission over land and over water, between night effects and day effects, and between transmission from

[1] Proceedings of I. R. E., Oct., 1926.

horizontal and from vertical antennas. Fading, speech quality and noise are discussed. The results are briefly interpreted in terms of present day short wave theories.

An Introduction to Ultra-Violet Metallography.[2] FRANCIS F. LUCAS. This paper describes the ultra-violet microscope and the technique of its application to the study of metal surfaces. The ultra-violet microscope can be said to have lived up to expectations. Crisp brilliant images can be obtained which surpass in quality those obtainable with the apochromatic system. The potential resolving ability of the monochromats can be realized in practice and the practical application of the ultra-violet microscope should develop much new information. The ultra-violet microscope is the most complicated of any within the realm of technical or scientific microscopy. It requires a highly developed technique for its successful manipulation and the specimens must be prepared with great care. The ultra-violet equipment appears to have a potential resolving ability probably greater than twice that of the apochromatic system.

Portable Receiving Sets for Measuring Field Strengths.[3] AXEL G. JENSEN. Describes a measuring set involving the use of a current-dividing potentiometer accurate for frequencies up to about 1,500 kilocycles and having a field-strength range of about 20 to 200,000 microvolts per meter.

Thermionic and Adsorption Characteristics of Caesium on Tungsten and Oxidized Tungsten.[4] JOSEPH A. BECKER. Curves showing the logarithm of the electron current per cm^2 from tungsten and oxidized tungsten over a wide range of filament temperatures are given for several vapor pressures of caesium. At high temperatures the tungsten is covered only to a slight extent with adsorbed caesium. As the filament temperature is lowered more caesium is adsorbed. This lowers the electron work function and increases the emission many thousandfold. The process continues until a temperature is reached at which the tungsten is just covered with a monatomic layer when the work function has a minimum value. At still lower temperatures the surface is more than completely covered, the work function increases again, and the emission decreases rapidly.

[2] Presented before the American Institute of Mining and Metallurgical Engineers, New York, N. Y., February, 1926. Published as Pamphlet No. 1576-E, issued with *Mining and Metallurgy.*

[3] Proceedings I. R. E., page 333, June, 1926.

[4] *Physical Review*, Vol. 28, pp. 341–361, August, 1926.

The positive ion emission is constant while the temperature decreases from a high value to a low critical temperature. Here the ion emission drops suddenly while some caesium sticks to the filament. Further decreases in temperature are followed by increased adsorption and decreased ion emission. If the temperature is then increased in steps the ion current retraces its path. At an upper critical temperature, about 50° higher than the lower critical temperature, the filament cleans itself spontaneously, the caesium comes off as ions and registers as a sudden rush of current. At higher temperatures the ion current has its initial constant value which is limited by the arrival rate of caesium atoms. The critical temperatures are raised by an increase in the vapor pressure or by a decrease in the plate potential.

A method of determining the amount of adsorbed caesium is developed. At a sufficiently high filament temperature the surface is clean. At a sufficiently low temperature every atom that strikes the filament sticks to it, at least until the optimum activity is reached. The product of the arrival rate, which is given by the steady positive ion current, and the time to attain the optimum activity gives the number of caesium atoms at the optimum activity. At an intermediate temperature the surface is only partly covered. If the temperature is suddenly dropped, to a low value, it takes a shorter time to reach the optimum activity. From these times the amount of adsorbed caesium at various temperatures, plate potentials, and vapor pressures can be determined. At the optimum activity there are 3.7×10^{14} atoms of caesium on a cm^2 of tungsten. This is very nearly the same as the number of caesium atoms that could be packed in a single layer, but is considerably smaller than the number of caesium ions in such a layer.

The adsorption or evaporation characteristics are illustrated by curves. Caesium can evaporate either as ions or as atoms. The atomic rate depends only on the temperature and on Θ, the fraction of the surface covered with caesium. For a given temperature it increases very rapidly with Θ. The ions can permanently escape from the filament only if the potential is in the right direction. A typical isothermal ion rate curve increases rapidly with Θ, comes to a maximum when Θ is about .01, then decreases continuously for larger Θ. These curves explain all the observed phenomena of these adsorbed films. They show that while the ion work function increases as Θ increases, the work to remove an atom decreases with Θ. The ion work function for a given Θ can be decreased by increasing the potential gradient at the filament.

The Significance of Magnetostriction in Permalloy.[5] L. W. McKeehan. Magnetostriction in permalloy confirms qualitatively the existence of atomic magnetostriction as previously proposed, and the explanation, based thereon, for high magnetic permeability and low hysteresis in these alloys. The effect of tension upon magnetostriction suggests that orientation of the magnetic axes of iron and nickel atoms precisely like that due to the application of magnetic fields may be effected by mechanical stresses within the elastic limit. Acceptance of this view makes it possible to explain the large effects of tension upon magnetic hysteresis and the observed relation between the changes in electrical resistance produced by tension and by magnetization. The occurrence of a reversal of magnetostriction in a stretched wire containing 80 per cent nickel is covered by the same explanation. A connection between magnetic hysteresis and mechanical hysteresis is suggested and the molecular field postulated by Weiss is interpreted as the integrated effect of local mechanical stresses.

Magnetostriction in Permalloy.[6] L. W. McKeehan and P. P. Cioffi. The materials studied comprised iron, nickel, and permalloys containing 46, 64, 74, 78, 80, 84, and 89 per cent nickel. The method permitted simultaneous measurement of magnetization and magnetostriction in about 12 cm. at the middle of a 40 cm. wire, 1 mm. in diaeter, in an approximately uniform field not exceeding 40 gauss, and either with or without applied tension (within the elastic limit).

The magnetostriction was measured by a combination of a mechanical lever, an optical lever, a multiple slit and a photoelectric cell. The magnifying power of this combination, as used, was about 2×10^6, and magnetostrictive strains from 2×10^9 to 3×10^5 were detected and measured without changing the sensitivity.

The magnetostriction-magnetization curve has initial slope zero in all the cases studied. When the attainable field was sufficient for magnetic saturation the magnetostriction reached a limiting value. In iron there is evidence for the existence of a Villari reversal in fields too great to be attained in the apparatus. In nickel there is no sign of such reversal. In the permalloys with more than 81 per cent Ni the magnetostriction is a contraction. In the permalloys with less than 81 per cent Ni the magnetostriction is an expansion. The limiting values of megnetostriction, when plotted against chemical composition, fall on a smooth curve.

[5] *Physical Review*, Vol. 28, page 158, July, 1926.
[6] *Physical Review*, Vol. 28, page 146, July, 1926.

Tension increases magnetostrictive contraction and diminishes magnetostrictive expansion. It causes a reversal in the sign of magnetostriction in permalloy with 80 per cent Ni, a small contraction preceding the final small expansion.

Transmission Features of Transcontinental Telephony.[7] H. H. NANCE and O. B. JACOBS. In this paper, the various steps in the establishment of the existing network of transcontinental type circuits and the transmission design considerations are reviewed. The discussion covers the communication channels obtained from transcontinental type facilities and the bands of frequencies employed, and includes carrier-current systems, telephone repeaters and signalling systems. Mention is made of special uses of transcontinental telephone circuits, such as the transmission of program material for broadcasting and the transmission of pictures. Finally, the maintenance methods required to keep the system at full efficiency are outlined.

[7] Presented at the Pacific Coast Convention of the A. I. E. E., Salt Lake City, Utah, Sept. 6–9, 1926.

Contributors to this Issue

S. E. Anderson, B.S. in E. E., University of Michigan, 1919; Electrical Research Laboratory, General Motors Corporation, Detroit, 1919–20; Engineering Department, Western Electric Company, 1920–24; Bell Telephone Laboratories, 1925—. Mr. Anderson has been engaged in the development of carrier systems and radio receivers.

John R. Carson, B.S., Princeton, 1907; E.E., 1909; M.S., 1912; Research Department, Westinghouse Electric and Manufacturing Company, 1910–12; instructor in 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 many papers on theory of electric circuits and electric wave propagation.

E. T. Hoch, B.S., in Electrical Engineering, Case School of Applied Science, 1914; Western Electric Company, Manufacturing and Installation Department, 1914–15; Engineering Department, 1915–24; Bell Telephone Laboratories, Inc., 1925—.

F. C. Willis, B.Sc., London University, 1911; Engineering Department of the Western Electric Company, 1913–14; 1919–24; British Army, 1914–19; Bell Telephone Laboratories, 1925—. Mr. Willis was formerly engaged on design of radio apparatus and now on line and transmission testing apparatus.

L. E. Melhuish, B.Sc., Pennsylvania State College, 1919; Engineering Department of the Western Electric Company and Bell Telephone Laboratories, 1919—. Mr. Melhuish has been employed on apparatus design relating to telephone repeaters, loud speaking systems and associated apparatus.

Walter A. Shewhart, A.B., University of Illinois, 1913; A.M., 1914; Ph.D., University of California, 1917; Engineering Department, Western Electric Company, 1918–24; Bell Telephone Laboratories, Inc., 1925—. Mr. Shewhart is making a special study of the application of probability theories to inspection engineering.

FRANCES THORNDIKE, A.B., Vassar College, 1922; M.A., Columbia University, 1925, American Telephone and Telegraph Company, Department of Development and Research, 1922—.

S. D. WILBURN—Long Lines Department of the American Telephone and Telegraph Company, 1904–1926; Line Construction, 1904–1906; Line Maintenance, 1906–1909; Equipment Maintenance, 1909–1912; Special Contract Service and Plant Maintenance, 1912–1918; Telephone Transmission Maintenance, 1918–1922; Telegraph Engineering, 1922–1926.

H. W. HITCHCOCK, B.S., Pomona College, 1911; Cornell University, 1911–13; Engineering Department of American Telephone and Telegraph Company, 1913–19; Department of Development and Research, 1919–21; The Pacific Telephone and Telegraph Company, Engineering Department, 1921–24; Transmission and Protection Engineer, Southern California Telephone Company, 1924—. Mr. Hitchcock's work prior to 1921 was in connection with inductive interference and submarine cable transmission development problems; his later work has been in connection with the transmission activities of one of the operating companies.

Index to Volume V

A

Advances in Physics; The Atom Model, page 96. Ionization, page 463, *K. K. Darrow.*

Alkali Metal Photoelectric Cell, The, *Herbert E. Ives,* page 320.

Amplifiers, Load Carrying Capacity of, *F. C. Willis* and *L. E. Melhuish,* page 573.

Anderson, S. E., Radio Signaling System for the New York Police Department, page 529.

Antenna Arrays, Directive Diagrams of, *Ronald M. Foster,* page 292.

Applications of Poisson's Probability Summation, *Frances Thorndike,* page 604.

Atom Model, The; Some Contemporary Advances in Physics, *Karl K. Darrow,* page 96.

B

Beck, C. J., and *C. F. Sacia,* The Power of Fundamental Speech Sounds, page 393.

Bown, Ralph, De Loss K. Martin and *Ralph K. Potter,* Some Studies in Radio Broadcast Transmission, page 143.

Bridge Polar Duplex Telegraph Circuits, Line Current Regulation in, *S. D. Wilburn,* page 625.

C

Cables, Carrier-Current Communication on Submarine, *H. W. Hitchcock,* page 636.

Carrier-Current Communication on Submarine Cables, *H. W. Hitchcock,* page 636.

Carson, John R., Electric Circuit Theory and the Operational Calculus, page 50; page 336.

Carson, John R., Wave Propagation in Overhead Wires with Ground Return, page 539.

Circuits, Telegraph, Line Current Regulation in Bridge Polar Duplex, *S. D. Wilburn,* page 625.

Circuit Theory and the Operational Calculus, Electric, *John R. Carson,* page 50; page 336.

Circuits, Operation of Thermionic Vacuum Tube, *F. B. Llewellyn,* page 433.

Control Charts, Quality, *W. A. Shewhart,* page 593.

Correction of Data for Errors of Averages Obtained from Small Samples, *W. A. Shewhart,* page 308.

Correction of Data for Errors of Measurement, *W. A. Shewhart,* page 11.

D

Darrow, Karl K., Some Contemporary Advances in Physics—the Atom Model, page 96. Ionization, page 463.

Development and Application of Loading for Telephone Circuits, *Thomas Shaw* and *William Fondiller,* page 221.

Dielectric Constant of Sheet Insulating Materials, Electrode Effects in the Measurement of Power Factor and, *E. T. Hoch,* page 555.

Directive Diagrams of Antenna Arrays, *Ronald M. Foster,* page 292.

E

Electric Circuit Theory and the Operational Calculus, *J. R. Carson*, page 50; page 336.

Electrode Effects in the Measurement of Power Factor and Dielectric Constant of Sheet Insulating Materials, *E. T. Hoch*, page 555.

Errors of Averages Obtained from Small Samples, Correction of Data for, *W. A. Shewhart*, page 308.

Errors of Measurement, Correction of Data for, *W. A. Shewhart*, page 11.

Extraneous Interference on Submarine Telegraph Cables, *J. J. Gilbert*, page 404.

F

Fletcher, Harvey, Theory of the Howling Telephone with Experimental Confirmation, page 27.

Fondiller, William and *Thomas Shaw*, Development and Application of Loading for Telephone Circuits, page 221.

Foster, Ronald M., Directive Diagrams of Antenna Arrays, page 292.

Friis, H. T., A Static Recorder, page 282.

G

Gherardi, Bancroft and *Robert W. King*, Joseph Henry—the American Pioneer in Electrical Communication, page 1.

Gilbert, J. J., Extraneous Interference on Submarine Telegraph Cables, page 404.

Ground Return, Wave Propagation in Overhead Wires with, *John R. Carson*, page 539.

H

Harrison, H. C. and *J. P. Maxfield*, Methods of High Quality Recording and Reproducing of Music and Speech Based on Telephone Research, page 493.

Henry, Joseph—The American Pioneer in Electrical Communication, *Bancroft Gherardi* and *Robert W. King*, page 1.

Hitchcock, H. W., Carrier-Current Communication on Submarine Cables, page 636.

Hoch, E. T., Electrode Effects in the Measurement of Power Factor and Dielectric Constant of Sheet Insulating Materials, page 555.

Howling Telephone with Experimental Confirmation, Theory of the, *Harvey Fletcher*, page 27.

I

Insulating Materials, Electrode Effects in the Measurement of Power Factor and Dielectric Constant of Sheet, *E. T. Hoch*, page 555.

Ionization, Contemporary Advances in Physics, *Karl K. Darrow*, page 463.

Ives, Herbert E., The Alkali Metal Photoelectric Cell, page 320.

K

King, Robert W. and *Bancroft Gherardi*, Joseph Henry—the American Pioneer in Electrical Communication, page 1.

L

Line Current Regulation in Bridge Polar Duplex Telegraph Circuits, *S. D. Wilburn*, page 625.

Llewellyn, F. B., Operation of Thermionic Vacuum Tube Circuits, page 433.

Load Carrying Capacity of Amplifiers, *F. C. Willis* and *L. E. Melhuish*, page 573.

Loading for Telephone Circuits, Development and Application of, *Thomas Shaw* and *William Fondiller*, page 221.

M

Martin, De Loss K., Ralph Bown and *Ralph K. Potter*, Some Studies in Radio Broadcast Transmission, page 143.

Maxfield, J. P. and *H. C. Harrison*, Methods of High Quality Recording and Reproducing of Music and Speech Based on Telephone Research, page 493.

Measurement, Correction of Data for Errors of, *W. A. Shewhart*, page 11.

Measurement of Power Factor and Dielectric Constant of Sheet Insulating Materials, Electrode Effects in the, *E. T. Hoch*, page 555.

Melhuish, L. E. and *F. C. Willis*, Load Carrying Capacity of Amplifiers, page 573.

Methods of High Quality Recording and Reproducing of Music and Speech Based on Telephone Research, *J. P. Maxfield* and *H. C. Harrison*, page 493.

Music and Speech, Based on Telephone Research, Methods of High Quality Recording and Reproducing of, *J. P. Maxfield* and *H. C. Harrison*, page 493.

N

Neutralization of Telegraph Crossfire, *R. B. Shanck*, page 418.

O

Operation of Thermionic Vacuum Tube Circuits, *F. B. Llewellyn*, page 433.

Operational Calculus, Electric Circuit Theory and the, *J. R. Carson*, page 50; page 336.

P

Photoelectric Cell, The Alkali Metal, *Herbert E. Ives*, page 320.

Physics, Some Contemporary Advances in, The Atom Model, page 96. Ionization, page 463; *Karl K. Darrow*.

Poisson's Probability Summation, Applications of, *Frances Thorndike*, page 604.

Police Department, Radio Signaling System for the New York, *S. E. Anderson*, page 529.

Potter, Ralph K., Ralph Bown and *De Loss K. Martin*, Some Studies in Radio Broadcast Transmission, page 143.

Power Factor and Dielectric Constant of Sheet Insulating Materials, Electrode Effects in the Measurement of, *E. T. Hoch*, page 555.

Power of Fundamental Speech Sounds, The, *C. F. Sacia* and *C. J. Beck*, page 393.

Probability Summation, Applications of Poisson's, *Frances Thorndike*, page 604.

Q

Quality Control Charts, *W. A. Shewhart*, page 593.

R

Radio Broadcast Transmission, Some Studies in, *Ralph Bown, De Loss K. Martin* and *Ralph K. Potter*, page 143.

Radio Signalling System for the New York Police Department, *S. E. Anderson*, page 529.

Recording and Reproducing of Music and Speech Based on Telephone Research, Methods of High Quality, *J. P. Maxfield* and *H. C. Harrison*, page 493.

S

Sacia, C. F. and *C. J. Beck,* The Power of Fundamental Speech Sounds, page 393.
Samples, Correction of Data for Errors of Averages Obtained from Small, *W. A. Shewhart,* page 308.
Shanck, R. B., Neutralization of Telegraph Crossfire, page 418.
Shaw, Thomas and *William Fondiller,* Development and Application of Loading for Telephone Circuits, page 221.
Shewhart, W. A., Correction of Data for Errors of Averages Obtained from Small Samples, page 308.
Shewhart, W. A., Correction of Data for Errors of Measurement, page 11.
Shewhart, W. A., Quality Control Charts, page 593.
Speech Sounds, The Power of Fundamental, *C. J. Beck* and *C. F. Sacia,* page 393.
Static Recorder, A, *H. T. Friis,* page 282.
Submarine Cables, Carrier-Current Communication on, *H. W. Hitchcock,* page 636.
Submarine Telegraph Cables, Extraneous Interference on, *J. J. Gilbert,* page 404.

T

Telegraph Cables, Extraneous Interference on Submarine, *J. J. Gilbert,* page 404.
Telegraph Circuits, Line Current Regulation in Bridge Polar Duplex, *S. D. Wilburn,* page 625.
Telegraph Crossfire, Neutralization of, *R. B. Shanck,* page 418.
Telephone Circuits, Loading for, Development and Application of, *Thomas Shaw* and *William Fondiller,* page 221.
Telephone, Howling, Theory of the, with Experimental Confirmation, *Harvey Fletcher,* page 27.
Theory of the Howling Telephone with Experimental Confirmation, *Harvey Fletcher,* page 27.
Thermionic Vacuum Tube Circuits, Operation of, *F. B. Llewellyn,* page 433.
Thorndike, Frances, Applications of Poisson's Probability Summation, page 604.
Transmission, Some Studies in Radio Broadcast, *Ralph Bown, De Loss K. Martin* and *Ralph K. Potter,* page 143.

V

Vacuum Tube Circuits, Operation of Thermionic, *F. B. Llewellyn,* page 433.

W

Wave Propagation in Overhead Wires with Ground Return, *John R. Carson,* page 539.
Wilburn, S. D., Line Current Regulation in Bridge Polar Duplex Telegraph Circuits, page 625.
Willis, F. C. and *L. E. Melhuish,* Load Carrying Capacity of Amplifiers, page 573.